Extensions of Randomization-Based Methods for Causal Inference

Extensions of Randomization-Based Methods for Causal Inference

A dissertation presented

by

Joseph Jiazong Lee

to

The Department of Statistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Statistics

Harvard University

Cambridge, Massachusetts

March 2015

Dissertation Advisor: Donald B. Rubin Joseph Jiazong Lee

Extensions of Randomization-Based Methods for Causal Inference

Abstract

In randomized experiments, the random assignment of units to treatment groups justifies many of the traditional analysis methods for evaluating causal effects. Specifying subgroups of units for further examination after observing outcomes, however, may partially nullify any advantages of randomized assignment when data are analyzed naively. Some previous statistical literature has treated all post-hoc analyses homogeneously as entirely invalid and thus uninterpretable. Alternative analysis methods and the extent of the validity of such analyses remain largely unstudied. Here Chapter 1 proposes a novel, randomization-based method that generates valid post-hoc subgroup p-values, provided we know exactly how the subgroups were constructed. If we do not know the exact subgrouping procedure, our method may still place helpful bounds on the significance level of estimated effects. Chapter 2 extends the proposed methodology to generate valid posterior predictive p-values for partially post-hoc subgroup analyses, i.e., analyses that compare existing experimental data — from which a subgroup specification is derived — to new, subgroup-only data. Both chapters are motivated by pharmaceutical examples in which subgroup analyses played pivotal and controversial roles. Chapter 3 extends our randomization-based methodology to more general randomized experiments with multiple testing and nuisance unknowns. The results are valid familywise tests that are doubly advantageous, in terms of statistical power, over traditional methods. We apply our methods to data from the United States Job Training Partnership Act (JTPA) Study, where our analyses lead to different conclusions regarding the significance of estimated JTPA effects. In all chapters, we investigate the operating characteristics and demonstrate the advantages of our methods through series of simulations.

Contents

Acknowledgements

I owe a deep and sincere thank you to:

- Professor Donald B. Rubin, for inspiring my passion for statistics and for devoting countless hours to discussing ideas and proofreading my work over the years.
- Luci Yang, Benjamin J. Lee, and my parents, for giving me their unwavering love and support.
- Professors Tirthankar Dasgupta, Luke Miratrix, and Natesh S. Pillai, for their collaboration and guidance.
- My cousins and friends, for making me smile.
- God, for making all things possible.

Chapter 1

Randomization-Based Inference for Post-Hoc Subgroups

Lee and Rubin (accepted, 2015a)

1.1 Introduction

A causal effect is a comparison of potential outcomes for a common set of units. The fundamental problem of causal inference is that at most one of the potential outcomes for each unit can ever be observed (Rubin 1974). Statisticians generally consider randomized experiments to be the gold standard for evaluating causal effects. In such instances, the random assignment of experimental units to treatment groups justifies many of the widely-used traditional analysis methods.

In addition to drawing conclusions about the entire collection of units, it is often desirable to consider subsamples of the dataset at hand. When considering multiple individually valid subgroup inferences, however, one must be aware of multiple comparisons, a well-known but complicated issue studied by numerous statisticians, including Miller (1981) and Tukey (1991). Traditional multiple comparisons adjustments (e.g., Bonferroni corrections) are often severely conservative when subgroups overlap (see Chapter 3 and Westfall and Young (1989)). Moreover, specifying subgroups of units after observing outcomes may make

it difficult to identify the number of comparisons being made, making such adjustments less straightforward.

Furthermore, post-hoc decisions may partially nullify any advantage of randomized assignment when data are analyzed naively. As a simple example, consider a randomized experiment with both female and male units. Suppose we observe no significant (at some level) treatment effect over the whole study population but observe that the treatment appears significantly beneficial for females. If only then do we specify females as the subgroup of interest, the traditionally calculated female p-value loses its traditional interpretation. When examining post-hoc subgroup inferences, it is natural to expect overstatements of statistical significance. Yet, there is a statistical sense of "more valid" or "less valid" that is ignored when all post-hoc analyses are homogenized as entirely invalid.

The issues associated with post-hoc subgroup inferences are not only theoretically challenging and interesting to statisticians, but also directly applicable to many fields of study, including medicine and the social sciences (Assmann et al. 2000; Peck 2003; Rothwell 2005; Wang et al. 2007). Post-hoc analyses can draw attention because of their evocative nature, and they often provide evidence for possible future studies. Follow-up experiments, however, are often associated with large monetary costs and long delays before new data can be obtained. Quick results and valid inferences from the dataset at hand, therefore, can be beneficial to both drug developers and their patients, especially when dealing with rare diseases. As opined by the clinical researchers of our motivating example, however, there are currently no established guidelines for evaluating post-hoc subgroup inferences (Raghu et al. 2004). Of course, the desire for timely analyses and valid inferences is not limited to clinical trials.

Section 1.2 describes a recent pharmaceutical case with First Amendment legal implications, in which post-hoc subgroup inferences played a pivotal and controversial role. Here we propose two approaches for obtaining valid p-values for such subgroups. First, in Section 1.3, we formulate a simulation model for the pharmaceutical experiment and use it to i)

demonstrate the operating characteristics of post-hoc estimates of subgroup average treatment effects (SubATEs) and subgroup p-values, and ii) make model-based adjustments to the subgroup p-values for more accurate interpretations. Second, in Section 1.4, we describe a novel extension of a well-established statistical procedure that generates statistically valid subgroup p-values, even when those subgroups are specified *a posteriori*, provided we know exactly what procedure was used to construct those subgroups; if we do not know the exact procedure, this method can still place helpful bounds on the significance level of estimated effects. Although we do not have the raw data for the example described in Section 1.2, we believe this approach could have an important impact on such examples and lead to more accurate judgments about subgroup causal effects than the overly crude, dichotomous classification of inferences as valid or invalid.

1.2 An Illustration

Idiopathic pulmonary fibrosis (IPF), a rare disease characterized by progressive scarring of the lung tissue, currently affects nearly 100,000 people in the United States (Cleveland Clinic 2013; InterMune 2013). In addition to eventual respiratory failure, IPF is associated with pulmonary hypertension (high blood pressure in the lungs), pulmonary embolism (blood clots in the lungs), lung infections, heart attack, stroke, and lung cancer. As such, IPF is debilitating and often fatal, with an estimated median survival time of two to three years after diagnosis (Raghu et al. 2004). According to the United States National Institutes of Health, there is no known cause of IPF (hence the name, "idiopathic"); as of this chapter's initial submission, no medication had yet been approved to treat it (A.D.A.M., Inc. 2013).

In 1999, InterMune began marketing the drug interferon gamma-1b, branded under the name Actimmune, which was approved in 2000 by the United States Food and Drug Administration (FDA) for the treatment of two rare diseases unrelated to IPF. Nevertheless, the drug could be, and was, prescribed off-label by pulmonologists to treat IPF (Stretch et al.

2010), and InterMune conducted a series of clinical trials designed to achieve a label change stating that Actimmune was effective for treating IPF.

However, the company's randomized Phase III trial (GIPF-001) failed to meet its primary endpoint regarding "progression-free survival" for its 330 IPF patients. The trial also failed to meet any of its nine secondary endpoints, one of which was survival time (Stretch et al. 2010). Although the (two-sided) survival time p-value of 0.084 was insignificant based on the 330 patients (the standard FDA cutoff for p-value significance is 0.05), the Actimmune group exhibited a 40% lower mortality rate than the control group. Furthermore, in the subgroup of 254 patients with mild to moderate disease (defined by baseline predicted forced vital lung capacity (FVC)), the Actimmune group exhibited a 70% lower mortality rate than the control group. Depending on the FVC criterion, the p-value for the specified subgroup of patients was as small as 0.004 (InterMune 2012). For any disease, especially one with consequences as grave as IPF, a real survival benefit is of the utmost interest, and indications of a survival benefit with such strong magnitude demand attention.

On August 28, 2002, under the direction of then-President and CEO W. Scott Harkonen, InterMune issued a press release titled, "InterMune Announces Phase III Data Demonstrating Survival Benefit of Actimmune in IPF: Reduces Mortality by 70% in Patients With Mild to Moderate Disease." The press release stated that the GIPF-001 data "...demonstrate a significant survival benefit in patients with mild to moderate disease randomly assigned to Actimmune versus the control treatment ($p = 0.004$)" (InterMune 2012, p. 1). In a medical sense, it may be reasonable to consider this subgroup of patients as highly relevant because the disease in the excluded patients (classified as having severe disease) may have already progressed beyond the point where any treatment could be effective. Nonetheless, as noted in Section 1.1, statistical issues arise whenever subgrouping procedures are performed, especially when they are performed after outcome data are observed.

Following the press release, the United States Department of Justice prosecuted Harkonen for "...fraudulently promoting the drug Actimmune" (*United States v. Harkonen* 2013),

citing the statistical invalidity of post-hoc subgroup analyses. The prosecution argued that Harkonen issued "... false and misleading information about the drug's effectiveness in treating idiopathic pulmonary fibrosis" in order to increase InterMune's revenue streams. On September 29, 2009, a jury found Harkonen guilty of wire fraud; in March 2013, Harkonen's conviction was affirmed by the Ninth Circuit of the United States Court of Appeals in *United States v. Harkonen*. A petition for writ of certiorari was denied by the U.S. Supreme Court in December 2013. For more legal context and discussion, see Brown (September 23, 2013). Legal arguments aside, the Actimmune case poses relevant and intriguing statistical questions about the validity of Harkonen's conclusions and of more general conclusions like them. Are post-hoc subgroup inferences interpretable in any way, and to what extent should they be trusted? What alternative analysis methods, if any, can be performed? Surely there must be something more that statisticians can do than assert "invalidity."

1.3 Actimmune Case Study: Simulation Model

In this section, we formulate a simulation model of the Actimmune case based on court documents, InterMune's internal documents, and extended conversations with Harkonen's legal team and Harkonen himself. Our goal is to assess, under the null hypothesis of zero treatment effect, how the presumably-used post-hoc subgrouping procedure distorts the traditionally calculated subgroup p-value.

1.3.1 Randomized Experiment

Suppose we have a randomized experiment with units indexed by $i = 1, ..., N_{total}$ and D binary covariates: $X_{id} = 1$ with probability p_d, $d = 1, ..., D$, where $\mathbf{p}_x = (p_1, ..., p_D)$ denotes the vector of probabilities. Further, suppose that N_1 units are randomly assigned to active treatment, and $N_0 = N_{total} - N_1$ are randomly assigned to control. The binary experimental outcome on which the treatment's effectiveness is to be assessed is Y. Each unit has two

potential outcomes: one under control, $Y_i(0)$, and one under the active treatment, $Y_i(1)$; at most one of these can ever be observed (Rubin 1974, 2005). This notation is sufficient under the stable unit treatment value assumption, which asserts no interference between experimental units, as well as two well-defined outcomes (Rubin 1980, 1986). Under the sharp null hypothesis of zero treatment effect, $Y_i(0) = Y_i(1)$ for all i.

For our simulation, we draw a "successful" potential outcome $Y_i(\cdot)$ with probability

$$P(Y_i(\cdot) = 1|X_{i1}, ..., X_{iD}, \beta_0, \beta_1, ..., \beta_D) = \min(1, \exp(\beta_0 + \beta_1 X_{i1} + ... + \beta_D X_{iD})),$$

where β_0 is the baseline log rate of success and $\beta_1, ..., \beta_D$ generate pseudo-correlations of $X_1, ..., X_D$ with Y. A positive β_d value suggests that covariate d increases survival probability multiplicatively when the other covariates are held constant. We select this model because it facilitates straightforward parameter estimation for the simulation setup from the available GIPF-001 data summaries; we check the model's fit in Section 1.3.2. Formally, the success probability of Y is truncated at 1, although boundary violations are not a concern within our simulation.

We estimate the following parameter values for the simulation using the available Actimmune data summaries; these values are then regarded as "truth" throughout the simulation:

- $N_{total} = 330, N_1 = 162, N_0 = 168$. (In reality, some early dropouts were recorded; these patients are ignored, as they were in the actual GIPF-001 trial analysis.)
- $D = 4$. The (pre-specified) relevant baseline covariates and their p_d vales are:
  - X_1: Baseline predicted forced vital capacity (FVC) ($\leq 60\%$ vs. $> 60\%$), $p_1 = 0.55$
  - X_2: Use of prednisone or equivalent at study entry (no vs. yes), $p_2 = 0.76$
  - X_3: Ratio of Actimmune dosage to body surface area (BSA) ($\leq 100\ \mu g/m^2$ vs. $> 100\ \mu g/m^2$), $p_3 = 0.48$
  - X_4: Days since IPF diagnosis (≤ 300 days vs. > 300 days), $p_4 = 0.51$
- $Y(0), Y(1)$ are survival indicators under the control and active treatments, respectively.

- The β values are:
  - Baseline FVC: $\beta_1 = 0.132$
  - Prednisone equivalent at baseline: $\beta_2 = -0.046$
  - Dosage-to-BSA ratio: $\beta_3 = -0.112$
  - Days since IPF diagnosis: $\beta_4 = 0.052$
- The aggregate survival rate (across all patients) was 86.7%. Considering p_d and β_d as known ($d = 1, ..., 4$), we then have four equations to obtain β_0: for $d = 1, ..., 4$,

$$\text{Overall survival probability} = p_d \exp(\beta_0 + \beta_d) + (1 - p_d) \exp(\beta_0) \approx 86.7\%.$$

  Using the mean of the four values of β_0, we set $\beta_0 = -0.147$.

1.3.2 Model Checks

To check the fit of our simulation model, we compare the key summary statistics of the GIPF-001 data to the empirical distributions based on 10,000 random datasets generated by our model (see Figure 1.1). The fit is excellent, both for the four covariates and for the aggregate survival percentage, suggesting that our simulation model realistically reflects the Actimmune scenario under the null hypothesis.

1.3.3 Post-Hoc Subgrouping Procedure

We outline Stages 0 and 1 of the Actimmune post-hoc subgrouping procedure as follows:

1. For each of the $D = 4$ covariates, the units are divided into two subgroups: those satisfying $X_{id} = 0$ and those satisfying $X_{id} = 1$, thereby creating eight subgroups. Each subject is in exactly four of the eight subgroups.
2. After observing outcome data, the researcher performs a traditional statistical test on the entire dataset, resulting in a Stage 0 estimated average treatment effect (ATE)

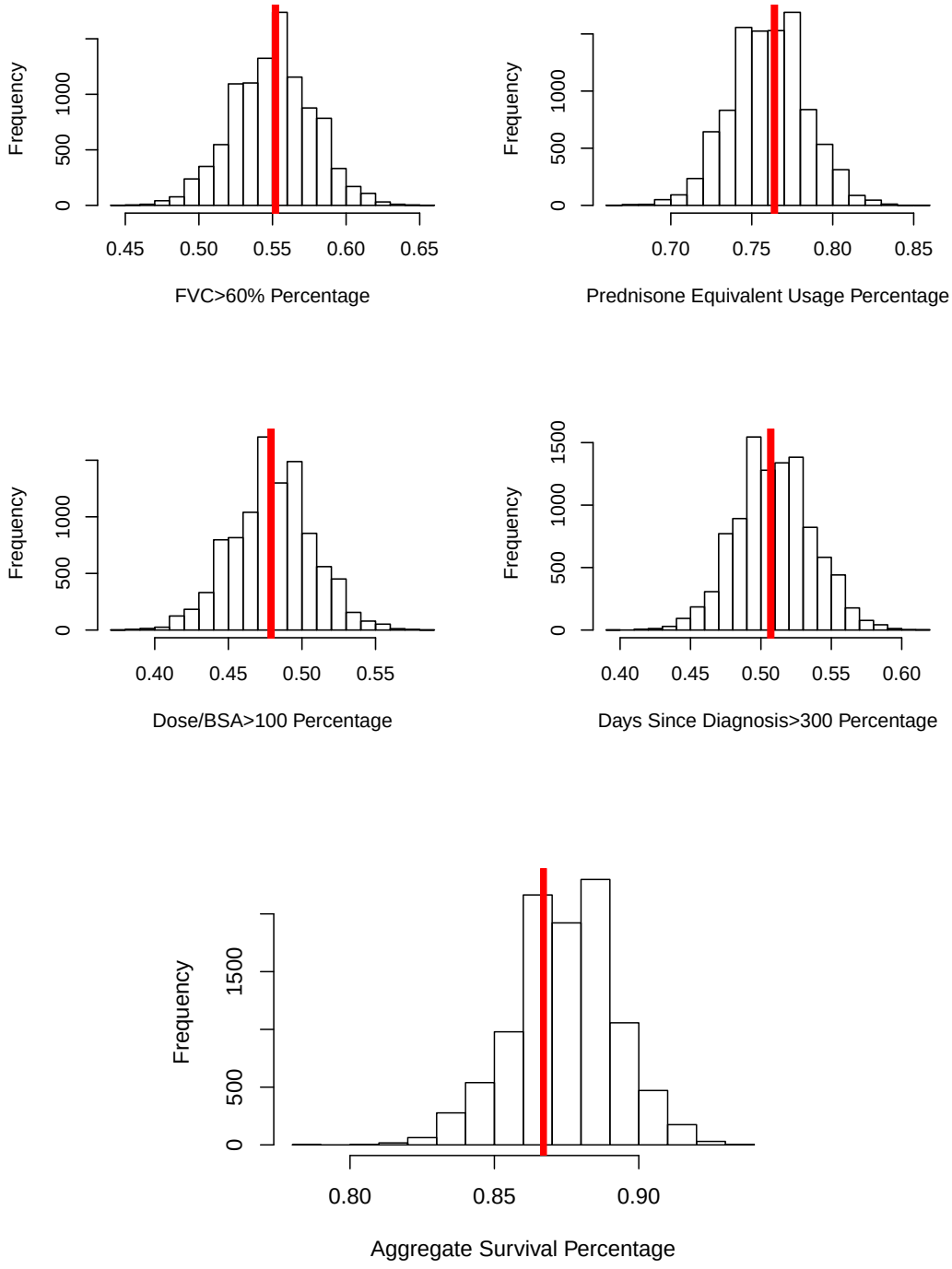

Figure 1.1: Comparison of actual observed covariate proportions and aggregate survival proportion to simulation model-generated empirical distributions. Observed proportions are represented by bold vertical lines. Empirical distributions are based on 10,000 model-generated datasets.

and corresponding p-value. Additionally, the researcher runs tests for each subgroup to estimate eight SubATEs and their eight corresponding p-values. We use a two-sample test for proportions, and the estimated ATE and SubATEs are calculated as differences in survival proportions. Nine tests are conducted in total.

3. The researcher, presumably wanting to describe the group or subgroup for which the treatment is most effective, selects — for Stage 1 reporting — the group or subgroup that shows the most evidence of a positive treatment effect, defined by the smallest p-value. If multiple subgroups share the minimum p-value, the largest subgroup (with respect to the number of included units) that shares that minimum p-value is selected for reporting. (If there are multiple largest subgroups that share the minimum p-value, one of the eligible subgroups is selected for reporting at random.) If the analysis on the entire dataset provides the smallest p-value, the Stage 0 p-value is reported.

In the Actimmune case, the Stage 1 subgroup included patients with baseline predicted FVC greater than 60%. Stage 2 involves further modification of this subgroup. According to court testimony by the GIPF-001 trial's biostatistician, InterMune researchers followed their Stage 1 analysis of FVC $\leq 60\%$ and $> 60\%$ subgroups with additional analyses of FVC $< 55\%$, $55 - 70\%$, and $> 70\%$ subgroups, before eventually defining the final Stage 2 subgroup of FVC $\geq 55\%$. Suppose $X_{(1)}$ is chosen as the Stage 1 covariate. We reflect the Stage 2 process in our simulation model as follows:

4. The $X_{(1)}$ categories from Stage 1 are further divided into two subcategories each, according to a probability vector, resulting in four ordered subcategories of covariate $X_{(1)}$. The four Actimmune subcategories were FVC: $< 55\%$, $55 - 60\%$, $60 - 70\%$, and $> 70\%$. We use the probability vector $(0.5, 0.5)$ because 36 of the 72 patients with FVC $\leq 60\%$ had FVC values below 55%, and we have no additional information about the patients with FVC $> 60\%$.

5. The nine possible Stage 2 subgroups are defined by unions of one or more adjacent subcategories. For example, possible Stage 2 Actimmune subgroups included FVC $< 55\%$, FVC $\in \{55-60\%,\ 60-70\%\}$ (i.e., $55\% \leq$ FVC $\leq 70\%$), and FVC $\in \{55-60\%, 60-70\%, > 70\%\}$ (i.e., FVC $\geq 55\%$). Similar to Stage 1, the researcher runs a traditional statistical test to estimate nine SubATEs and their nine corresponding p-values.

6. The final Stage 2 subgroup is the subgroup that shows the most significant subgroup p-value. The Actimmune Stage 2 subgroup was FVC $\in \{55-60\%, 60-70\%, > 70\%\}$, i.e., FVC $\geq 55\%$.

1.3.4 Simulation Results and Case Discussion

We generate 10,000 random datasets according to the null model described in Section 1.3.1 and perform Stage 1 and Stage 2 subgroup selection and analysis for each. Stage 0 tests (performed on the entire dataset, before post-hoc subgroup selection) behave conventionally under the null hypothesis of zero treatment effect: estimated ATEs are centered at zero and p-values are close to uniformly distributed, allowing for valid, traditional significance testing of estimated causal effects. After Stage 1, however, as expected, estimated SubATEs are biased upward, and subgroup p-values are skewed toward zero. Simulation results and comparisons are displayed in Table 1.1 and Figure 1.2.

	Subgroup Traditional p-value					
	Observed	Empirical Model Quantiles				
	Actimmune	5%	25%	50%	75%	95%
Stage 0	0.08	0.05	0.25	0.50	0.75	0.95
Stage 1	0.02	0.01	0.05	0.14	0.29	0.58
Stage 2	0.004	0.004	0.03	0.08	0.17	0.39

Table 1.1: Comparison of observed Actimmune p-values to empirical quantiles of Stage 0, Stage 1, and Stage 2 p-values based on 10,000 simulated replications under the null hypothesis.

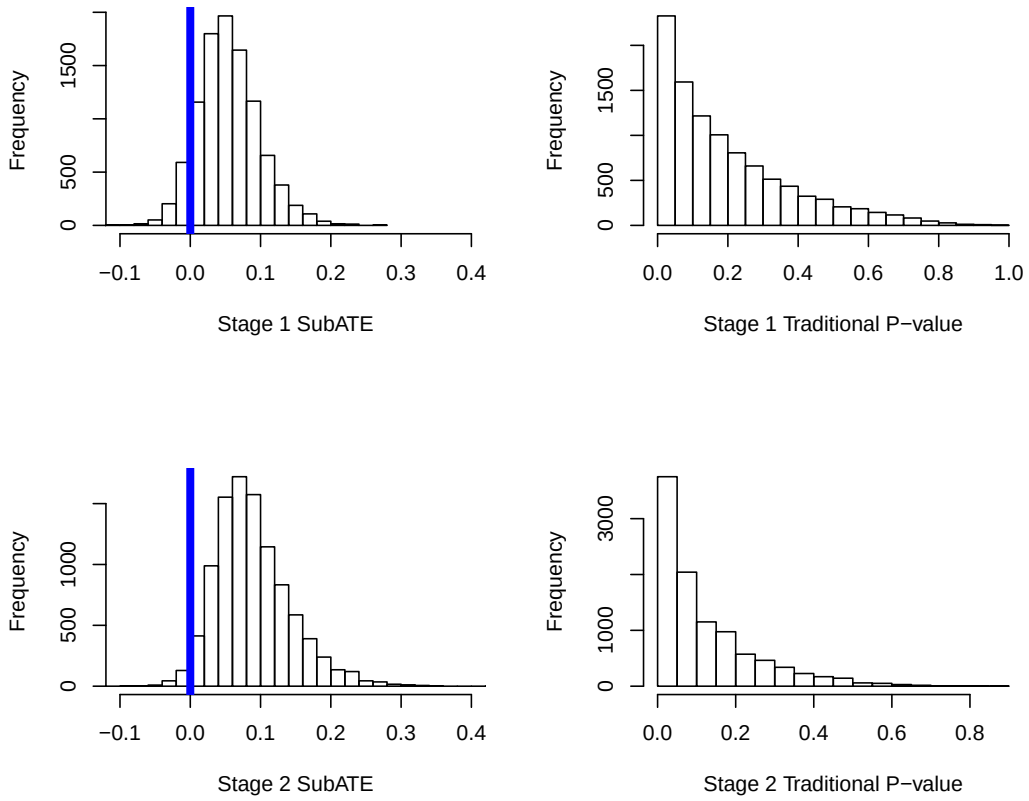

Figure 1.2: Empirical distribution of Stage 1 and Stage 2 results based on 10,000 simulated replications under the null hypothesis. Estimated SubATEs in both stages are biased upward rather than centered at zero. Bold vertical lines are placed at zero to highlight the bias. Subgroup p-values in both stages are skewed toward zero, rather than uniformly distributed, inflating the apparent significance of the post-hoc subgroup analyses.

Naturally, Stage 2 estimated SubATEs are further biased upward from Stage 1, with subgroup p-values further skewed toward zero. However, although the additional subgroup modification in Stage 2 often inflates significance further than Stage 1, the effect is not dramatic. In fact, as shown in the right panel of Figure 1.3, Stage 1 and Stage 2 subgroup p-values are actually identical in over 30% of simulated replications. On the whole, the simulation results illustrate that even when no treatment effect truly exists, Stage 1 and Stage 2 subgroup analyses often falsely suggest that the active treatment has a significant causal effect on the outcome for the chosen subgroup.

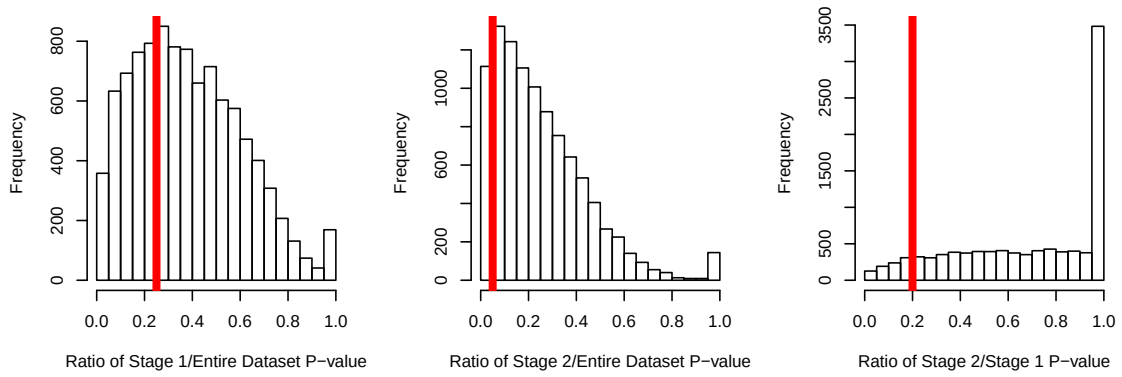

Figure 1.3: Empirical distribution of p-value ratios based on 10,000 simulated replications under the null hypothesis, with observed ratios from the GIPF-001 trial marked by bold vertical lines. Left: ratio of Stage 1 to Stage 0 p-values. Middle: ratio of Stage 2 to Stage 0 p-values. Right: ratio of Stage 2 to Stage 1 p-values.

The Actimmune Stage 0 p-value for survival was 0.08 — a p-value that may give some indication of a clinical benefit, although not technically "statistically significant" at the pre-specified 0.05 level. The Stage 1 and Stage 2 p-values from the Actimmune trial were 0.02 and 0.004, respectively. These values represent small fractions of the Stage 0 p-value (25% and 5%, respectively) and would both be considered highly significant at the 0.05 level *if* they had resulted from one analysis specified *a priori*. In order to interpret the *a posteriori* subgroup p-values more properly, we evaluate the frequency of such p-values under the null hypothesis, according to our simulation model.

After accounting for the post-hoc subgrouping procedure described in Section 1.3.3, both Stage 1 and Stage 2 p-values make dramatic movements toward insignificance. Specifically, based on the simulations under the null hypothesis, the Stage 1 p-value of 0.02 is even less significant than the Stage 0 p-value, moving from the calculated 0.02 to 0.11, as suggested by Figure 1.2. The Stage 2 p-value moves from the calculated 0.004 to 0.044; although the Stage 2 p-value technically remains "statistically significant" at the 0.05 level, its interpretation as overwhelming evidence of the drug's survival benefit no longer applies. To assess the inflation of statistical significance by stage, we also evaluate the frequency of p-value ratios of Stage

1 and 2 p-values to Stage 0 according to our simulation model. Under the null hypothesis, such extreme ratios of inflation are moderately, but not extremely, unlikely, occurring in about 32% and 11% of simulated replications, respectively (see Figure 1.3). Stage 1 to Stage 2 inflations as extreme as Actimmune's occur in about 9% of simulated replications. A complete evaluation of frequency characteristics is displayed in Table 1.2.

	Observed	Simulation Model Under Null Hypothesis
p-value	Actimmune	Proportion As Extreme As Observed
Stage 0	0.08	8.0%
Stage 1	0.02	11.2%
Stage 2	0.004	4.4%
	Observed	Simulation Model Under Null Hypothesis
p-value Ratio	Actimmune	Proportion As Extreme As Observed
Stage 1 to 0	0.25	32.5%
Stage 2 to 0	0.05	11.2%
Stage 2 to 1	0.2	8.7%

Table 1.2: Observed Actimmune test statistics and corresponding proportions of simulated replications for which significance or significance inflation was as extreme. Based on 10,000 simulated replications under the null hypothesis.

Because the simulation-based Stage 2 p-value is near the border of 0.05-level significance, we can surmise that accounting for more extreme subgrouping procedures would indeed push this p-value into traditional insignificance, tempering the apparent evidence of a treatment effect. More extreme procedures in the context of the Actimmune may also be more realistic. Namely, recall that mortality was one of nine secondary endpoints in the GIPF-001 trial; on the other hand, the subgrouping procedure in Section 1.3.3, for simplicity, treated mortality as the sole outcome of interest. Post-hoc subgrouping procedures that incorporate the primary endpoint, the eight other secondary endpoints, or possibly additional covariates would very likely push the Stage 2 p-value to insignificance. (In fact, a follow-up study to the GIPF-001 trial showed no survival benefit from Actimmune over a placebo for a somewhat similar subgroup of patients (King Jr et al. 2009).)

1.4 A Randomization-Based Approach for Generating Valid Post-Hoc Subgroup p-values

A customized simulation model, such as the Actimmune one described in Section 1.3, will always require certain assumptions and may not always be plausible. We now propose a randomization-based approach that *does* generate valid p-values, provided we know exactly how the subgroups were constructed. If we do not know exactly how they were constructed, the method can still generate helpful bounds on the significance level associated with estimated causal effects, using reasonable approximations of the post-hoc procedure.

The distributions of the estimated ATE and of the associated p-value from the experiment can be viewed under the sharp null hypothesis of no treatment effect as functions solely of the random assignment vector. As an extension, the distributions of any estimated SubATE and traditionally calculated subgroup p-value can also be viewed as functions solely of the assignment vector, when the subgrouping procedures are known. Thus, by generating a random assignment vector according to the assignment mechanism, we can generate an estimated SubATE and a corresponding subgroup p-value for the post-hoc subgrouping procedure under this sharp null hypothesis. Over repeated realizations of the assignment vector, we can obtain the null randomization distributions of these subgroup statistics and thereby evaluate the significance of the original, observed subgroup statistics.

The basic form of this approach, in which no post-hoc subgrouping occurs, was first proposed by Fisher (1935) and has gained popularity through the Fisher exact test. The exact test is typically only used in situations with small sample sizes because the traditional normality-based asymptotic tests are often sufficiently accurate for situations with even moderate sample sizes, and computation of randomization-based tests for larger samples has been computationally demanding until recently. However, it is clear that for post-hoc subgroup analyses, the traditional tests are no longer interpretable in the usual fashion. Additionally, increased computational power has made larger-scale randomization tests relatively

manageable (for example, see Lee, Dasgupta, and Rubin (submitted) and the R package **randomizationInference** (Lee and Dasgupta 2013–2015)).

Our main contribution is to re-frame the observed post-hoc procedure as just one path, randomly realized from a larger decision tree. In other words, our solution considers what subgrouping and inferential steps would have been taken if the data had realized under a different randomization. Our method is similar in spirit to the randomization-based methods described in Brown and Fears (1981) and Westfall and Young (1989) for pre-specified multiple comparisons of multivariate binomial outcomes. In fact, the methods coincide when the post-hoc inference is the result of a simple set of multiple comparisons; in that setting, Westfall and Young (1989) showed that such randomization-based methods are more powerful than traditional multiple comparisons adjustments (e.g., Bonferroni). Our method, by contrast, applies more broadly because we do not assume an *a priori* fixed set of groups being examined or a specific type of experimental outcome (e.g., binomial). Namely, our method is capable of handling complicated post-hoc inferential procedures that may involve non-standard test statistics — such as the sequential selection procedure outlined in Section 1.3, whose examination set varies with the observed outcome data. By outputting a "true randomization-based p-value" for the observed dataset, the proposed test allows us to assess the significance of estimated causal effects from post-hoc subgroups in a valid manner, without any additional model assumptions and without relying on asymptotic approximations. We see the proposed modified randomization test as a compelling and straightforward extension of Fisher's ideas, much like the re-randomization p-values discussed in Morgan and Rubin (2012). We believe this randomization-based approach could have an important impact and could help, in some cases, alleviate the need for costly and time-consuming follow-up studies.

The modified randomization test for valid post-hoc subgroup p-values is as follows:

1. Specify precisely the post-hoc subgrouping procedure (e.g., Stage 1 and 2 Actimmune procedures) and the subgroup test statistic of interest, T (e.g., Stage 2 p-value).

2. Perform the post-hoc subgrouping procedure on the observed dataset to obtain the observed subgroup test statistic, T^{obs}.

3. Impute the missing potential outcomes under the sharp null hypothesis.

4. Draw a random hypothetical assignment vector according to the actual assignment mechanism. Treating the hypothetical assignment vector as true, create the corresponding hypothetical observed dataset using the complete set of observed and imputed potential outcomes.

5. Perform the post-hoc subgrouping procedure and calculate T on the hypothetical observed data to obtain T^{hyp}. Record whether this statistic is as extreme as T^{obs}.

6. Repeat Steps 4–5 a reasonable number of times. If sample size is small, it may be possible to cycle through all possible assignment vectors. In most cases, however, drawing a large number (e.g., 10,000, depending on computational constraints) of random assignment vectors, with replacement, provides a sufficient approximation.

7. Calculate the proportion of hypothetical observed datasets for which T^{hyp} is as extreme as or more extreme than T^{obs}. This is the "true randomization-based" p-value (or Fisher p-value). If all possible permutations are used, this proportion is the exact Fisher p-value.

In some applied settings, it may be difficult to specify the post-hoc procedure exactly as it occurred. In such cases, our method may still place helpful bounds on the significance level of estimated effects, using reasonable approximations of the post-hoc procedure. By specifying a "conservative" version of the post-hoc procedure that results in the same observed test statistic, we can establish an upper bound for the true p-value of the estimated effect. We classify a version of the procedure as "conservative" if it leads to more significant estimated effects, in expectation, than the actual procedure. If this upper bound is already below the pre-specified significance cutoff, then we are assured that the true p-value is also below the

cutoff. Similarly, specifying a "liberal" version of the post-hoc procedure — one that leads to less significant estimated effects in expectation — can establish a lower bound for the true p-value. If the lower bound is already above the pre-specified cutoff, then we are assured that the true p-value is also above the cutoff.

To demonstrate our method's properties, we apply the proposed modified randomization test to 1,000 randomly generated datasets from the Actimmune simulation model, with the Stage 2 p-value as the subgroup test statistic. For each dataset, we draw 100 random hypothetical assignment vectors (chosen for computational convenience) and calculate a randomization-based p-value. As shown in Figure 1.4, the randomization-based p-value achieves the desired property of being approximately uniformly distributed under the null hypothesis. Given the actual GIPF-001 trial data, this randomization-based approach could help us provide a more definitive judgment about the statistical significance of the treatment effect on the FVC $\geq 55\%$ subgroup. Because we do not have the raw dataset, however, it is difficult to make such a judgment without utilizing the simulation model from Section 1.3.

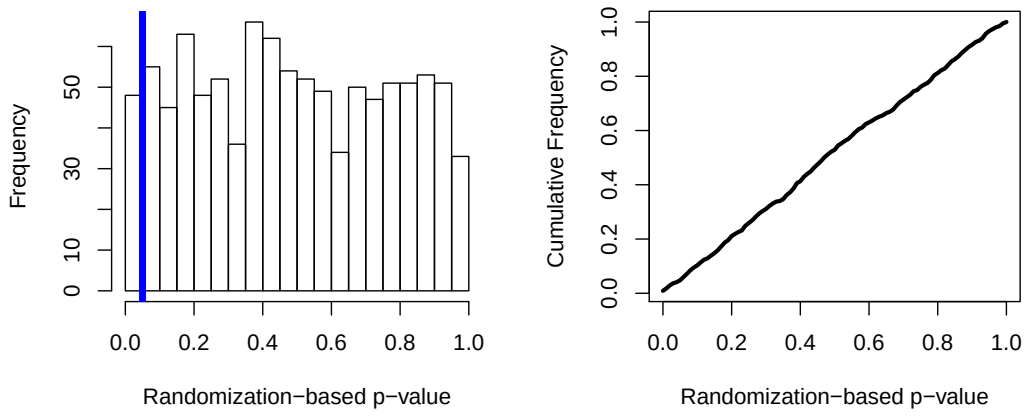

Figure 1.4: Empirical distribution of the randomization-based p-value based on 1,000 randomly generated datasets under the null hypothesis. For each dataset, the randomization-based p-value is calculated using 100 random draws of the assignment vector. The desired property of approximate uniformity is achieved, with approximately 5% of p-values below 0.05 (denoted by the bold vertical line).

This randomization-based method makes no distributional assumptions about the test statistic and can thus be particularly useful when the test statistic is unconventional, as seen in the Actimmune case. The method, however, is not limited to assessing estimated causal effects in post-hoc subgroups; in fact, it is valid for assessing causal effects under very general settings, including, but not limited to, situations examining multiple outcomes of interest, differences in outcome spreads, differences in outcome ranks, and subgroups resulting from pre-specified decision trees.

1.5 Conclusion

The severity of diseases like IPF demands valid and swift evaluations of drug efficacy, but the Actimmune case is just one of many real-world examples in which post-hoc subgroup inferences have played an important role. Due to the undeniable popularity of exploratory post-hoc analyses and the understandable desire to avoid additional data collection costs, the statistically valid consideration of these issues deserves investigation. We believe the proposed randomization-based method could have an important impact on post-hoc subgroup analyses, thereby leading to more accurate and more efficient judgments about subgroup causal effects.

Chapter 2

Randomization-Based Inference for Partially Post-Hoc Subgroups

Lee and Rubin (accepted, 2015b)

2.1 Introduction

Subgroup causal effects are often of scientific interest in randomized experiments. When subgroups are specified after observing outcomes, however, the estimated subgroup effects and p-values produced by traditional statistical methods do not have the typically desired repeated sampling properties, as described in Chapter 1. Traditional multiple comparisons (e.g., Bonferroni) adjustments tend to be overly conservative when subgroups overlap (see Chapter 3 and Westfall and Young (1989)). Moreover, post-hoc decisions often make it difficult to specify the number of comparisons being made, making such adjustments less straightforward. "Partially post-hoc" subgroup analyses, which compare existing data — from which the subgroup specification is derived — to new, subgroup-only experimental data, are further complicated.

Here we describe a motivating example faced by the U.S. FDA in which a partially post-hoc subgroup analysis instigated statistical debate about a medical device's efficacy. We provide a statistical framework to clarify the source of statistical invalidity. We then

propose a randomization-based method for generating valid posterior predictive p-values for such partially post-hoc subgroups. Although we do not have raw data for the particular example, we investigate the method's operating characteristics through a series of simulations, showing that it exhibits both a valid type I error rate and substantial power under reasonable alternative hypotheses.

2.2 The Durolane Trials

In March 2006, Swedish medical device company Q-Med AB submitted its medical device Durolane to the FDA for pre-market approval. Durolane is a viscous gel intended to treat osteoarthritic knee pain, administered through intra-articular (into-the-joint) injection. Prior to FDA review, the device had already been approved in a number of countries across Europe and Asia. As evidence of Durolane's efficacy, Q-Med submitted analyses of data from three randomized clinical trials.

Q-Med initially conducted two randomized, double-blind experiments attempting to demonstrate Durolane's superiority to a saline placebo (control), measuring each patient's pre- and post-treatment pain scores on the Western Ontario and McMasters Universities Osteoarthritis Index (WOMAC). Higher WOMAC scores translate to higher levels of pain; patients were deemed positive responders to treatment if their WOMAC pain scores were reduced by at least 40% and at least 5 points on the 0–20 point scale. Comparisons of responder rates in Studies 1 and 2 produced p-values of 0.53 and 0.49, respectively, neither of which are considered close to statistically significant by the FDA (the standard FDA cutoff for significance is 0.05).

After observing and un-blinding outcome data from Studies 1 and 2, Q-Med combined and filtered the data for a post-hoc subgroup analysis. Patients without effusion (fluid accumulation in the knee joint) and without baseline polyarticular pain — 284 (50.4%) of the 564 total patients — were included in the selected subgroup. Researchers asserted that

including patients with effusion and polyarticular pain would make it "... difficult to observe a treatment effect because the presence of these conditions leads to considerable variability in the WOMAC assessment" (Q-Med, AB 2009). The subgroup, however, was not specified *a priori*. Within the subgroup, Q-Med found a statistically significant difference in responder rates, boasting a drastically decreased p-value of 0.0013.

Per FDA's request, Q-Med conducted a third clinical trial to confirm this affirmative result for the specified patient subpopulation. All Study 3 patients satisfied the covariate inclusion criteria specified in the post-hoc analysis of Studies 1 and 2. Instead of using a placebo control group in Study 3, however, Q-Med decided to compare Durolane to the drug methylprednisolone in a non-inferiority trial. Although methylprednisolone is currently an approved standard of care for osteoarthritic knee pain, there exist concerns about its side effects, particularly its tendency to destroy cartilage over time; Durolane is purportedly able to avoid this detrimental side effect. Based on 442 patients, Study 3 successfully showed Durolane's non-inferiority to methylprednisolone. But because no saline control group was included, Study 3 alone did not provide direct evidence, as desired by the FDA, for Durolane's superiority to a saline placebo in reducing knee pain.

Because there was no placebo control group for Study 3, Q-Med used covariate-matched saline placebo patients from Studies 1 and 2 to assess Durolane's effectiveness. The selected historical placebo controls (i) met Study 3 covariate inclusion criteria, and (ii) provided sufficient covariate balance in comparison to the Study 3 Durolane treatment group, as determined by a propensity score model (Rosenbaum and Rubin 1983). Observed outcomes were not used in the selection of historical controls. The comparison of historical controls from Studies 1 and 2 to Durolane patients from Study 3 favored Durolane, with a statistically significant p-value of 0.047. Nevertheless, in August 2009, the FDA rejected Durolane for sale in the U.S.

2.3 A Statistical Framework

Here we describe a statistical framework and explore the implications of using historical controls in the Durolane context. We argue why the combined subgroup p-value is invalid (in terms of type I error) and demonstrate that balancing covariate profiles (e.g., via propensity score matching) cannot fully repair its validity.

2.3.1 Experiment 1 and Post-hoc Subgroup Specification

Suppose we have N_1 patients in Experiment 1 (representing Durolane Studies 1 and 2 collectively), each defined by a single, binary covariate X (e.g., male/female) and randomly assigned to the control or active treatment (indicated by W). Let Y be a binary experimental outcome representing whether or not a patient "responds" to the assigned treatment (e.g., in terms of having reduced knee pain). The "Science" table (Rubin 2005) for Experiment 1 and its observed values under a particular assignment are shown in Table 2.1. Each patient has two potential outcomes (Neyman 1923), only one of which can ever be observed (Rubin 1974). This notation is sufficient under the stable unit treatment value assumption, which asserts no interference between experimental units, as well as two well-defined outcomes (Rubin 1980). Unit i, j represents the jth unit from Experiment i.

Experiment 1						
	Covariate	Potential Outcomes		Assignment	Observed Outcomes	
Unit (i,j)	$X_{i,j}$	$Y_{i,j}(0)$	$Y_{i,j}(1)$	$W_{i,j}^{\text{obs}}$	$Y_{i,j}(0)$	$Y_{i,j}(1)$
$1,1$	0	$Y_{1,1}(0)$	$Y_{1,1}(1)$	0	$Y_{1,1}^{\text{obs}}$	?
$1,2$	1	$Y_{1,2}(0)$	$Y_{1,2}(1)$	0	$Y_{1,2}^{\text{obs}}$	?
$1,3$	0	$Y_{1,3}(0)$	$Y_{1,3}(1)$	1	?	$Y_{1,3}^{\text{obs}}$
...		...			...	
$1,N_1$	1	$Y_{1,N_1}(0)$	$Y_{1,N_1}(1)$	1	?	Y_{1,N_1}^{obs}

Table 2.1: The Science table for Experiment 1 (left) its and corresponding observed values under a particular assignment (right).

Suppose that after Experiment 1 outcomes are observed, the researcher calculates subgroup p-values for each possibly scientifically relevant subgroup, defined by X values. In

the setting with a single binary covariate X, there are three possible subgroups: $X = 0$, $X = 1$, and $X \in \{0, 1\}$. The researcher then presumably identifies the subgroup S for which the active treatment appears to have the most beneficial effect with respect to Y, e.g., the subgroup with the most significant p-value in favor of the active treatment.

2.3.2 Experiment 2

Experiment 2 (representing Durolane Study 3) consists of N_2 patients, all of whom satisfy the subgroup criteria of S and are assigned to active treatment; there are no control patients. In reality, Durolane's Experiment 2 introduced a new, third treatment condition (methylprednisolone, see Section 2.2). However, Q-Med's ultimate goal was to satisfy FDA's request to compare Durolane to the saline placebo (the control from Experiment 1). The Experiment 2 patients assigned to methylprednisolone are not relevant for this purpose and are thus ignored here.

Without loss of generality, suppose that the $X = 1$, e.g., female, subgroup is chosen from Experiment 1. Then Experiment 2 consists entirely of female patients assigned to the active treatment; the observed values of the Experiment 2 Science table are shown in Table 2.2.

Experiment 2				
	Covariate	Assignment	Observed Outcomes	
Unit (i, j)	$X_{i,j}$	$W_{i,j}^{\text{obs}}$	$Y_{i,j}(0)$	$Y_{i,j}(1)$
$2, 1$	1	1	?	$Y_{2,1}^{\text{obs}}$
$2, 2$	1	1	?	$Y_{2,2}^{\text{obs}}$
$2, 3$	1	1	?	$Y_{2,3}^{\text{obs}}$
...		...		
$2, N_2$	1	1	?	Y_{2,N_2}^{obs}

Table 2.2: The observed values of the Science table for Experiment 2.

2.3.3 Combined Subgroup Analysis

Because Experiment 2 outcomes are realized after S is specified, Experiment 2 data can be used to estimate the subgroup average treatment effect (SubATE) validly through traditional

statistical methods. For instance, if control patients were also included in Experiment 2, the data from Experiment 2 alone could generate a valid subgroup p-value. However, comparisons using Experiment 1 control patients and Experiment 2 treatment patients cannot be handled in the same manner.

Consider the following two quantities: (i) $\hat{\tau}_{1,S} = \bar{Y}^{\text{obs}}_{1,j\in S}(1) - \bar{Y}^{\text{obs}}_{1,j\in S}(0)$, and (ii) $\hat{\tau}_{\text{combined},S} = \bar{Y}^{\text{obs}}_{2,j\in S}(1) - \bar{Y}^{\text{obs}}_{1,j\in S}(0)$. The first quantity, $\hat{\tau}_{1,S}$, represents the estimated SubATE for S from Experiment 1, from which the selected subgroup specification is derived. The second, $\hat{\tau}_{\text{combined},S}$, represents the estimated SubATE for S obtained by comparing Experiment 1 control patients and Experiment 2 treatment patients. The combined subgroup analysis calculates $\hat{\tau}_{\text{combined},S}$ and generates an associated subgroup p-value.

2.3.4 Invalidity Under the Null Hypothesis

Suppose that the null hypothesis of zero treatment effect is true. Because the specification of S is a function of observed Experiment 1 outcomes, the expectation of $\hat{\tau}_{1,S}$ (over the randomization) is positive, and its associated p-value is skewed right, making traditional testing invalid in terms of type I error (see Chapter 1). Conceptually, $E(\hat{\tau}_{1,S}) > 0$ because $\hat{\tau}_{1,S}$ is the maximum or near-maximum of several estimated SubATEs from Experiment 1. Examining the two terms that comprise $\hat{\tau}_{1,S}$, we expect $\bar{Y}^{\text{obs}}_{1,j\in S}(1)$ to be high and $\bar{Y}^{\text{obs}}_{1,j\in S}(0)$ to be low because of the subgroup specification; in other words, we expect Experiment 1 treatment patients in S to have good outcomes, and Experiment 1 control patients in S to have poor outcomes, because such outcome information was used to select S in the first place.

In addition, because $\hat{\tau}_{\text{combined},S}$ shares one term with $\hat{\tau}_{1,S}$, we can see by the linearity of the expectation operator that $\hat{\tau}_{\text{combined},S}$ also has a positive expectation and a skewed-right p-value under the null hypothesis. Although $\bar{Y}^{\text{obs}}_{2,j\in S}(1)$ is realized after S is specified, the carry-over usage of $\bar{Y}^{\text{obs}}_{1,j\in S}(0)$ renders traditional testing of $\hat{\tau}_{\text{combined},S}$ invalid (though, one could argue, "less invalid" than testing of $\hat{\tau}_{1,S}$).

Finally, note in this setting that the historical controls in the combined subgroup analysis have covariate profiles (e.g., $X = 1$) that exactly match the Experiment 2 treatment patients. Here the statistical problem is rooted not in any discrepancies between control and treatment covariate profiles, but in the usage of observed Experiment 1 outcomes under the false assumption that they are independent of the subgroup specification. Any traditionally-calculated subgroup p-value for $\hat{\tau}_{\text{combined},S}$ cannot be valid, regardless of any covariate balancing (e.g., propensity score matching) techniques designed to mitigate that invalidity.

2.4 Valid Randomization-based p-values for Post-hoc Subgroups in the Presence of Nuisance Unknowns

Chapter 1 introduced a randomization-based approach for generating valid post-hoc subgroup p-values, motivated by earlier ideas about randomization due to Fisher (1935). The fundamental insight is to specify the decision tree that led to the final test statistic value, considering what subgrouping and inferential steps would have been taken if the data had been realized under a different randomization. The approach entails (i) specifying a precise post-hoc subgrouping procedure and an accompanying test statistic, (ii) calculating the test statistic on the observed data, (iii) imputing the missing potential outcomes in the study under a sharp null hypothesis, (iv) repeatedly drawing random hypothetical assignments according to the assignment mechanism and calculating test statistic values on the corresponding hypothetical datasets to construct the null randomization distribution of the test statistic, and (v) comparing the observed test statistic value against its null randomization distribution.

Calculating test statistic values on hypothetical data from a single experiment under a sharp null hypothesis is straightforward. For example, under the sharp null hypothesis of zero treatment effect, the missing potential outcomes can be imputed exactly as observed for

each unit. The test statistic is then calculated using the hypothetical assignment and corresponding hypothetical observed data, given the specified subgrouping procedure. However, in settings with multiple experiments — including one or more that occur after the subgroup specification — constructing the null randomization distribution of the test statistic requires some ingenuity.

Consider our example from Section 2.3, in which females ($X = 1$ patients) comprise the selected subgroup, S. The final test statistic, $\hat{\tau}_{\text{combined},S}$, compares outcomes from female Experiment 1 control patients with outcomes from Experiment 2 treatment patients, where by construction, all of the Experiment 2 patients are female. But what if males had exhibited a more beneficial estimated treatment effect than females in Experiment 1? Presumably, Experiment 2 would then have included only males; the experimental sample for Experiment 2 would have been completely different.

Here we propose an extension of the aforementioned method that generates valid posterior predictive p-values (Rubin 1984; Meng 1994) in the presence of nuisance unknowns, e.g., male Experiment 2 outcomes. We expand the Experiment 2 Science table, conceptualizing it as an augmented experiment with N_2' patients ($N_2' > N_2$), containing patients with the same mix of X values as Experiment 1 and filled with missing data. For Experiment 2 patients that exist in reality (i.e., females), potential outcomes under active treatment are observed, but potential outcomes under control are missing (unobserved). For Experiment 2 patients that exist only in our augmented framework (i.e., males), both potential outcomes are missing. The observed and unobserved values of the augmented Experiment 2 Science table are shown in Table 2.3.

Because the male treatment potential outcomes are not observed in Experiment 2, values of $\hat{\tau}_{\text{combined},S}$ under hypothetical randomizations can be considered random variables, with uncertainty resulting from these missing potential outcomes. Given a set of imputed values, however, we can construct the randomization distribution of $\hat{\tau}_{\text{combined},S}$ and calculate a randomization-based p-value. Thus, by multiply imputing (Rubin 1987) the missing male

Experiment 2				
	Covariate	Assignment	Potential Outcomes	
Unit (i,j)	$X_{i,j}$	$W_{i,j}^{\text{obs}}$	$Y_{i,j}(0)$	$Y_{i,j}(1)$
$2,1$	1	1	?	$Y_{2,1}^{\text{obs}}$
$2,2$	1	1	?	$Y_{2,2}^{\text{obs}}$
$2,3$	1	1	?	$Y_{2,3}^{\text{obs}}$
...		...		
$2,N_2$	1	1	?	Y_{2,N_2}^{obs}
$2,N_2+1$	0	1	?	?
$2,N_2+2$	0	1	?	?
...		...		
$2,N_2'$	0	1	?	?

Table 2.3: The observed and unobserved values of the augmented Experiment 2 Science table.

potential outcomes according to a distributional model that assumes the null hypothesis, they can be "integrated out" to produce a posterior predictive p-value; the posterior predictive p-value is the average p-value over the multiple imputations of the missing treatment potential outcomes.

Under the null hypothesis, the posterior predictive distribution of the missing treatment potential outcomes is informed by the observed Experiment 1 potential outcomes for both control and treatment patients. In the framework with binary outcomes and two independent covariate subgroups (males versus females), there are two parameters to model for imputation: the probability of a successful male outcome and the probability of a successful female outcome. A typical Bayesian model involves two independent Beta priors and Binomial likelihoods, leading to a Beta-Binomial posterior predictive distribution for the missing treatment potential outcomes. Modeling is further simplified by the fact that the treatment potential outcomes in Experiment 2 need to be imputed only for the male patients; the parameter governing the female potential outcomes can be ignored because it is not needed.

For a post-hoc subgroup test statistic T, the full procedure for obtaining a posterior predictive p-value in the two-experiment setting is as follows:

1. Specify precisely the post-hoc subgrouping procedure and the subgroup test statistic

of interest, T (e.g., $\hat{\tau}_{\text{combined},S}$).

2. Perform the post-hoc subgrouping procedure on the observed dataset to obtain the observed subgroup test statistic, T^{obs}.

3. Using the augmented Experiment 2 framework, multiply impute the missing treatment potential outcomes (e.g., M times), using their posterior predictive distributions according to a distributional model that assumes the null hypothesis. M is a large number (e.g., 10,000) that controls the Monte Carlo integration error.

4. For each of the M imputed datasets, calculate a randomization-based p-value for T according to the actual assignment mechanism(s) and specified subgrouping procedure (see Lee and Rubin accepted, 2015a), treating the imputed values as true. The randomization-based p-value is the proportion of hypothetical randomizations for which the corresponding test statistic value, T^{hyp}, is as extreme as or more extreme than T^{obs}.

5. The posterior predictive p-value for the null hypothesis with respect to T equals the average of the M randomization-based p-values.

The method outlined above can be viewed as a form of data augmentation (Tanner and Wong 1987), with the expanded Experiment 2 population making it possible to obtain a posterior predictive p-value for T. Rubin (1998) described a similar procedure in a different setting, in order to obtain a posterior predictive p-value for the complier average causal effect (CACE) in a single experiment with non-compliance. In that setting, the nuisance unknowns were the missing compliance statuses of the patients in the experiment who were assigned to the control treatment. In the same paper, a computational shortcut was identified: for each of the M imputed datasets, only one hypothetical assignment needs to be drawn in Step 4. The individual randomization-based p-values then equal either 0 or 1, and the posterior predictive p-value is the average of the indicators.

As mentioned in Chapter 1, specifying a post-hoc procedure exactly as it occurred may be difficult. In such cases, the randomization-based posterior predictive p-values can still place

helpful bounds on the significance level of estimated subgroup effects by using reasonable approximations that place limits on the post-hoc procedure.

2.5 Operating Characteristics

2.5.1 Simulation Setup

To evaluate the operating characteristics of the proposed method, we simulate random datasets under various treatment effect hypotheses. We first randomly sample $N_1 = 500$ patients for Experiment 1, drawing from a population of 50% females and 50% males. We randomly assign $N_1/2$ of these patients to control and the other $N_1/2$ to active treatment. We then draw random Bernoulli outcomes according to the probabilities in Table 2.4.

	Males		Females	
	Control	Active Treatment	Control	Active Treatment
Null hypothesis A	0.50	0.50	0.50	0.50
Null hypothesis B	0.20	0.20	0.80	0.80
Alternative hypothesis A	0.50	0.55	0.50	0.55
Alternative hypothesis B	0.20	0.30	0.80	0.90

Table 2.4: Simulated outcome success ("response") probabilities under various treatment effect hypotheses.

After observing Experiment 1 outcomes, we specify the subgroup S for further study. There are three possible choices for S: males, females, or all patients; S is the subgroup exhibiting the smallest p-value based on control versus treatment Experiment 1 responder rates. (If multiple subgroups share the smallest p-value, the one in that pool with the largest number of included units is selected. If multiple subgroups share the smallest p-value and the largest sample size within that pool, one of them is selected at random.)

Experiment 2 is conducted with new patients, all of whom satisfy the criterion of S and are assigned to active treatment. The combined subgroup p-value is calculated for $\hat{\tau}_{\text{combined},S}$, comparing Experiment 2 treatment units in S to Experiment 1 control units in

S. Randomization-based posterior predictive p-values are then generated according to the procedure described in Section 2.4.

2.5.2 Simulation Results

Under the null hypotheses described in Table 2.4, both the Experiment 1 and combined (Experiments 1 and 2) subgroup p-values are invalid in terms of type I error, i.e., both subgroup p-values incorrectly reject the null hypothesis more often than desired by the nominal significance level. Figure 2.1 shows the histograms of these p-values under the null hypotheses. As expected, all of the histograms are heavily skewed right, indicating the p-values' invalidity. The combined subgroup p-values are slightly less skewed, suggesting that they are, in some sense, "less invalid" than the Experiment 1 subgroup p-values.

On the other hand, the posterior predictive p-value appears valid — in fact, conservative — in terms of type I error (see Figure 2.2). In other words, when the null hypothesis is true, the posterior predictive p-value rejects it less often than indicated by the nominal significance level. Such conservatism often arises when multiply imputing missing data under a null hypothesis (see Rubin 1998) and seems to become more extreme when the proportion of missing data is large.

Table 2.5 displays simulation results, comparing the type I error rates based on the Experiment 1, combined, and posterior predictive subgroup p-values.

	Type I Error Rate at $\alpha = .05$	
Subgroup p-value	Null Hypothesis A	Null Hypothesis B
Experiment 1	10.3%	12.3%
Combined (Experiments 1 and 2)	8.7%	8.0%
Posterior predictive	2.0%	2.3%

Table 2.5: Type I error rates (at $\alpha = .05$) of Experiment 1, combined, and posterior predictive subgroup p-values. Based on 1,000 simulated datasets.

Conservatism under a null hypothesis is often welcome, especially in FDA contexts, provided the method exhibits sufficient power under reasonable alternative hypotheses; simula-

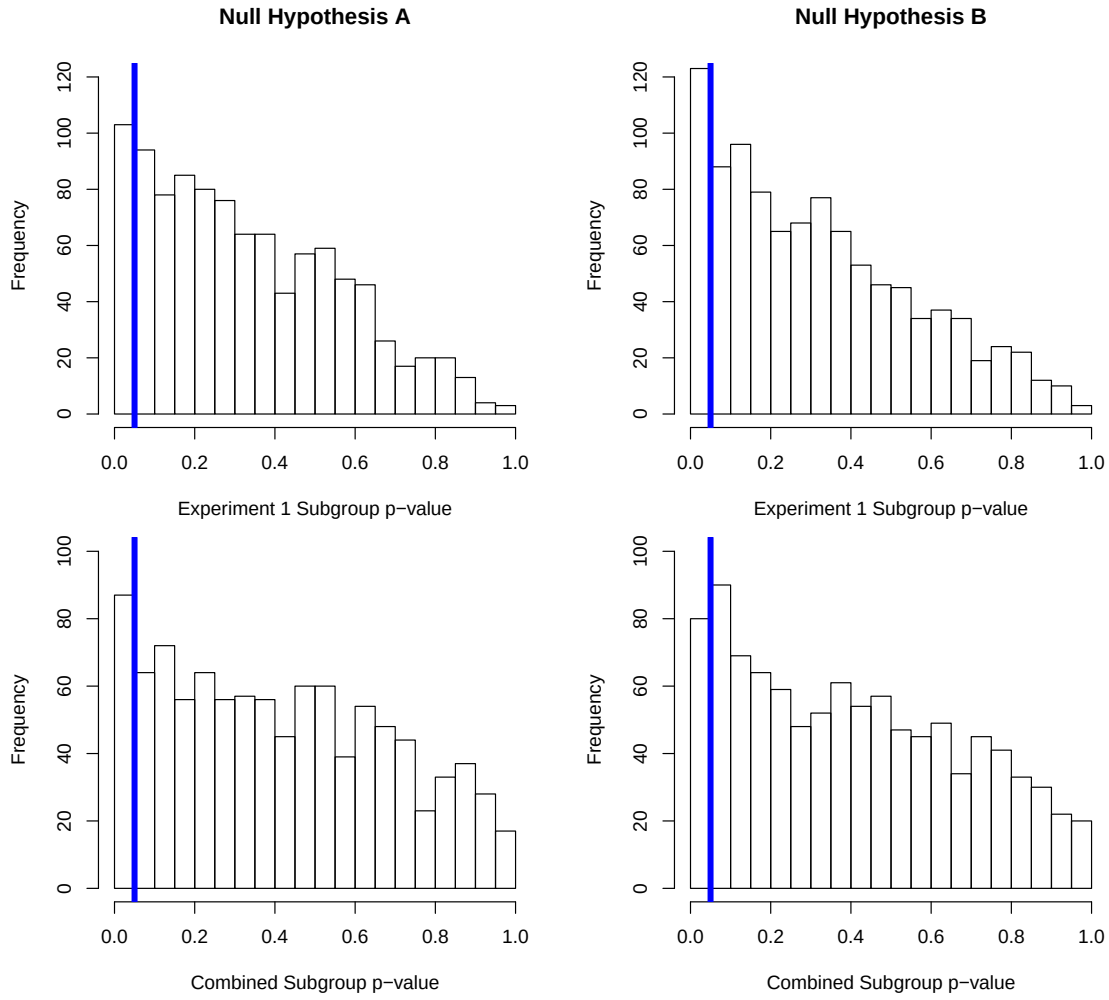

Figure 2.1: Empirical distributions of Experiment 1 (top) and combined (Experiments 1 and 2; bottom) subgroup p-values under the null hypotheses described in Table 2.4. Based on 1,000 simulated datasets.

tions show this to indeed be the case. Under the alternative hypotheses described in Table 2.4, the method has substantial power, rejecting the null hypothesis at $\alpha = .05$ in 21% and 69% of replications under Alternative Hypotheses A (5% treatment effect) and B (10% treatment effect), respectively. Figure 2.3 displays the histograms of posterior predictive p-values under these alternative hypotheses. We also note that in the motivating example from Section 2.2, pre-experiment sample size calculations aimed to capture 80% power at $\alpha = .05$ assuming larger treatment effects of 15–20%; according to simulations, our method

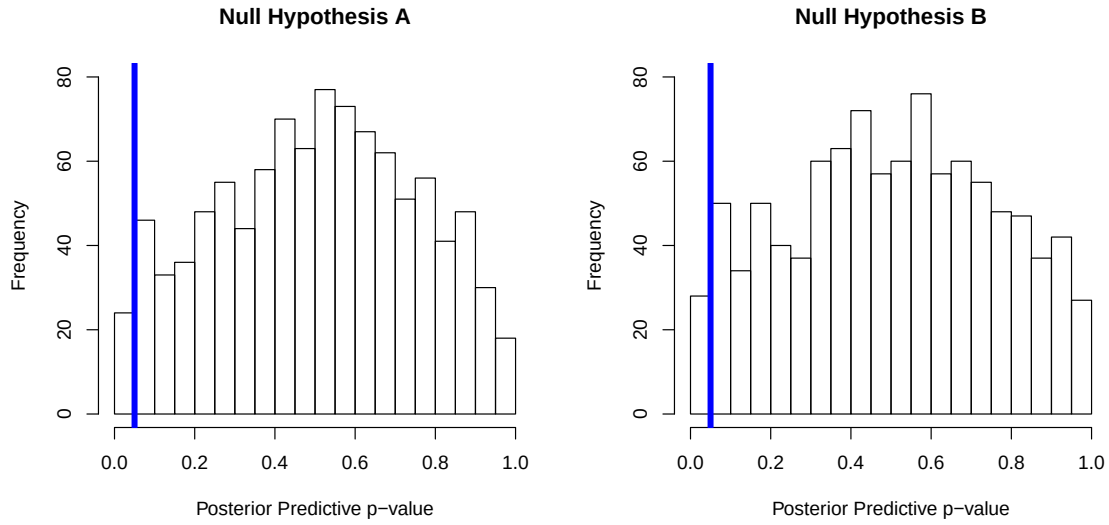

Figure 2.2: Empirical distributions of posterior predictive p-values under the null hypotheses described in Table 2.4. Based on 1,000 simulated datasets.

achieves over 95% power when generating data with such large effects.

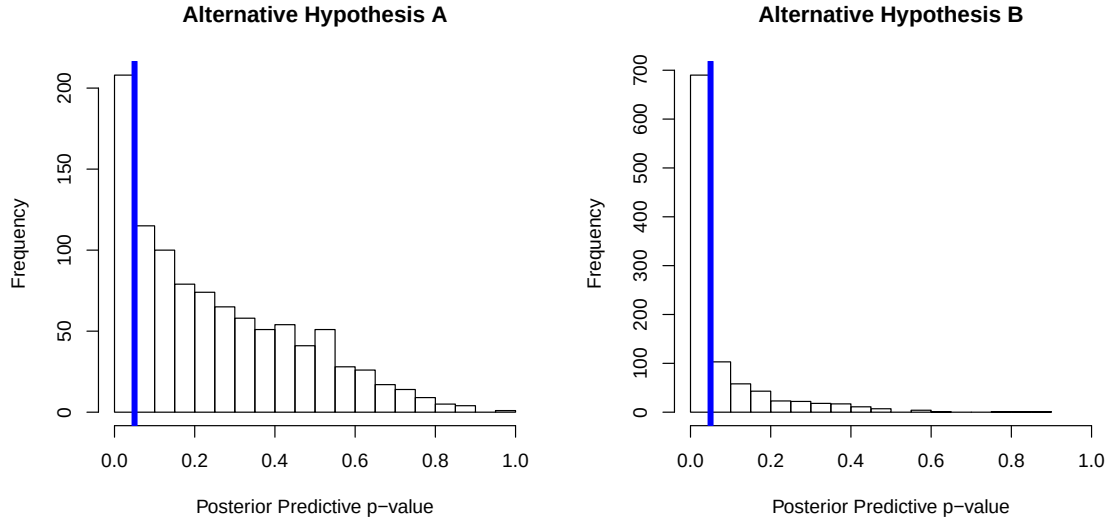

Figure 2.3: Empirical distributions of posterior predictive p-values under the alternative hypotheses described in Table 2.4. Based on 1,000 simulated datasets.

2.6 Conclusion

We have presented a randomization-based approach for generating valid posterior predictive p-values for partially post-hoc subgroups. We have also demonstrated that the resulting p-values have substantial power under reasonable alternative hypotheses. In the artificial example provided, the multiple imputation of the missing Experiment 2 treatment potential outcomes is facilitated by a conjugate model assuming independence between the subgroups. Such independence may not always be plausible. For instance, only two covariates — effusion and polyarticular pain — defined the Durolane subgroup; thus, subgroup patients could and did share other covariate values with non-subgroup patients. The augmented framework then requires an imputation model that relates covariates to the outcomes, e.g., Bayesian logistic regression, under the null hypothesis. The missing Experiment 2 treatment potential outcomes are again multiply-imputed according to their posterior predictive distribution, which can be empirically constructed using Markov Chain Monte Carlo techniques. More generally, our randomization-based approach applies to randomized experiments with nuisance unknowns; examples of such unknowns include missing compliance statuses (see Chapter 3) as well as missing outcome data from patients who exist in reality (as opposed to existing only in the augmented framework).

Chapter 3

More Powerful Multiple Testing in Randomized Experiments with Non-Compliance

Lee, Miratrix, and Pillai (submitted)

3.1 Introduction

The United States Job Training Partnership Act (JTPA) Study was a randomized experiment in the 1980s designed to measure the effects of a national, publicly-funded training program. Participants randomly assigned to the treatment group were eligible to receive JTPA services, while participants randomly assigned to the control group were barred from JTPA services for 18 months. Only about 2/3 of the treatment participants, however, actually enrolled and received any JTPA services; the other 1/3 failed to comply with their treatment assignment. Furthermore, because of the fluid nature of the participants' employment, researchers were interested in measuring JTPA effects across several time periods after random assignment, including the in-training period and the first and second post-program years. Analyzing such data requires addressing two substantial concerns: (i) due to non-compliance, the effects of treatment assignment are not equivalent to the effects of treatment

receipt, and (ii) conducting tests for multiple time periods without appropriate adjustments may lead to an inflated type I error rate. In this chapter, we outline an analysis method that addresses both concerns while maintaining reasonable power to detect treatment effects.

When units in randomized experiments fail to comply with their random assignment, inference for the effects of treatment receipt, rather than of assignment alone, becomes less straightforward. Intention-to-treat (ITT) analyses, which ignore treatment receipt, may have low power when assignment alone has no effect on the experimental outcome. In order to address this loss of power, Rubin (1998) introduced randomization-based posterior predictive p-values for the complier average causal effect (CACE) and showed through simulation that (i) they are valid p-values in terms of type I error, and (ii) they have higher power than ITT p-values under reasonable alternative hypotheses. This framework follows the general approach for Bayesian causal inference in randomized experiments with non-compliance outlined by Imbens and Rubin (1997). Both pieces of work rely on the multiple imputation (Rubin 1987) of missing compliance statuses; separating the experimental units into principal strata (Frangakis and Rubin 2002) based on compliance behavior aids inference for the desired causal effect. We use these tools in our approach but adapt them for simultaneous testing of multiple outcomes and subgroups.

Multiple testing issues are common in randomized experiments because multiple outcomes and subgroups of interest are often measured and analyzed for possible effects. Traditionally, practitioners have applied Bonferroni corrections to sets of p-values in order to control their familywise error rate (FWER), i.e., the rate at which at least one type I error is made, in a straightforward manner. Bonferroni corrections, however, tend to be overly conservative, especially when those p-values are correlated (Westfall and Young 1989). Such conservatism has led many applied researchers to avoid Bonferroni corrections and abandon multiple comparisons adjustments altogether (Cabin and Mitchell 2000; Nakagawa 2004; Perneger 1998; Rothman 1990). Other avenues exist; randomization-based procedures can provide greater power while maintaining the FWER by accounting for correlated tests. Brown

and Fears (1981) and Westfall and Young (1989) first introduced permutation-based multiple testing adjustments, though they did not explicitly motivate them using randomized assignment mechanisms. Randomization-based procedures are additionally appealing because they do not require any assumptions about the underlying distribution (here, joint) of the data. Furthermore, recent increases in computational power have helped such procedures become more tractable and gain popularity (Good 2005).

In this chapter, we connect and extend methodological ideas to appropriately handle both non-compliance and multiple testing in randomized experiments. We build up to this combined approach in stages. In Section 3.2, we elucidate the method proposed by Rubin (1998) for evaluating meaningful causal effects in the presence of non-compliance. In Section 3.3, we extend the ideas of Westfall and Young (1989) to fully randomization-based multiple comparisons adjustments and propose such adjustments as a straightforward yet more powerful alternative to Bonferroni corrections. In Section 3.4, we merge the notions of non-compliance and multiple testing, and outline a combined method of analysis that demonstrates power advantages from both perspectives. In each of Sections 3.2–3.4, we empirically show the benefits of the described methods through a series of simulated experiments. In Section 3.5, we apply traditional methods and our combined method to JTPA data to evaluate the program's effects on employment rate by time period. We illustrate how the methods lead to different conclusions regarding the significance of estimated JTPA effects. Section 3.6 concludes.

3.2 Experiments with Non-compliance

3.2.1 Non-compliance as a missing data problem

Suppose we have a randomized experiment with N units, indexed by i, with observed covariates X_i, randomly assigned to control or active treatment. Let Z_i be a binary indicator for assignment to active treatment, and let $D_i(z)$ be a binary indicator for receipt of active

treatment under assignment z. A unit's compliance behavior C_i is defined by the pair of potential outcomes (Neyman 1923; Rubin 1974) $(D_i(0), D_i(1))$; this notation is sufficient under the stable unit treatment value assumption (Rubin 1980, 1986), which asserts no interference between experimental units, as well as two well-defined outcomes. Each unit then belongs to one of four possible compliance strata:

- Compliers ($C_i = c$), who receive their treatment assignment: $(D_i(0), D_i(1)) = (0, 1)$.
- Never-takers ($C_i = nt$), who never receive the active treatment: $(D_i(0), D_i(1)) = (0, 0)$.
- Always-takers ($C_i = at$), who always receive the active treatment: $(D_i(0), D_i(1)) = (1, 1)$.
- Defiers ($C_i = d$), who receive the opposite of their treatment assignment: $(D_i(0), D_i(1)) = (1, 0)$.

If non-compliance is one-sided — i.e., units assigned to control are prohibited from receiving the active treatment — then $D_i(0) = 0$ for all i. In such settings, always-takers and defiers do not exist, and two possible strata are left: compliers and never-takers. Real-world scenarios involving one-sided non-compliance include many clinical trials, in which new drugs are unavailable to control patients, and some job training experiments, in which training programs and additional services are unavailable to the control group.

In many practical settings, researchers are most interested in the compliers because the effect of treatment assignment is synonymous with the effect of treatment receipt for those units. Strata membership, however, can never be fully determined for all units because they depend on the two potential outcomes of D, one of which is missing (i.e., unobserved). Membership can, on the other hand, be partially determined based on the observed potential outcome, D_i^{obs}. Table 3.1 outlines the possible compliance strata based on units' observed treatment assignment and receipt. An example "Science" table (Rubin 2005) under one-sided non-compliance and its observed values under a particular assignment are shown in Table 3.2.

Assignment	Receipt	Possible C_i Values	
Z_i	D_i^{obs}	One-sided Non-compliance	Two-sided Non-compliance
0	0	c, nt	c, nt
0	1	–	at, d
1	0	nt	nt, d
1	1	c	c, at

Table 3.1: Units' possible compliance strata based on observed treatment assignment and receipt.

		$D(z)$		Compliance	$Y(z)$		Assignment	$D(z)$		Compliance	$Y(z)$	
Unit	X_i	$D_i(0)$	$D_i(1)$	C_i	$Y(0)$	$Y(1)$	Z_i	$D_i(0)$	$D_i(1)$	C_i	$Y_i(0)$	$Y_i(1)$
1	X_1	0	0	nt	$Y_1(0)$	$Y_1(1)$	0	0	?	?	Y_1^{obs}	?
2	X_2	0	1	c	$Y_2(0)$	$Y_2(1)$	1	0	1	c	?	Y_2^{obs}
3	X_3	0	1	c	$Y_3(0)$	$Y_3(1)$	1	0	1	c	?	Y_3^{obs}
4	X_4	0	0	nt	$Y_4(0)$	$Y_4(1)$	1	0	0	nt	?	Y_4^{obs}
...				...						...		
N	X_N	0	1	c	$Y_N(0)$	$Y_N(1)$	0	0	?	?	Y_N^{obs}	?

Table 3.2: An example Science table under one-sided non-compliance (left) and its corresponding observed and unobserved values under a particular assignment (right).

Because strata memberships are not fully observed, uncertainty with respect to complier-specific effects stems from the missing compliance statuses (i.e., D potential outcomes) in addition to the missing Y potential outcomes. One approach to handling the additional uncertainty is to, in a Bayesian framework, view the missing compliance statuses as random variables. By multiply imputing the missing compliance statuses, e.g., according to a distributional model, they can be "integrated out," and we can make inference specific to the compliers.

3.2.2 Randomization-based posterior predictive p-values

As described by Meng (1994), a posterior predictive p-value can be viewed as the posterior mean of a classical p-value, averaging over the posterior distribution of nuisance factors (e.g., missing compliance statuses) under the null hypothesis. Rubin (1998) introduced a randomization-based procedure, which we expound on here, for obtaining posterior predictive p-values for estimated complier-only effects. One posterior predictive p-value is the

average of many p-values calculated from multiple "compliance-complete" datasets with imputed compliance statuses; for each compliance-complete dataset, the p-value is obtained through a randomization test (Fisher 1925, 1935). Within one randomization test, however, calculations of the test statistic do not use all of the compliance data; rather, they use only the compliance data that would have actually been observed under particular hypothetical randomizations. Though implied, this step of re-observing the data is not explicitly stated by Rubin (1998); we place it in Step 5 of the procedure below for emphasis.

In this section, we assume a single outcome for simplicity. The procedure for obtaining a randomization-based posterior predictive p-value is as follows:

1. **Choose a test statistic and calculate its observed value.**

   Choose a test statistic, T, to estimate an effect on the outcome, Y. Examples include the maximum-likelihood estimate (MLE) of CACE or the posterior median of CACE, given the observed compliance statuses and potential outcomes, under the exclusion restriction (see Angrist et al. 1996; Imbens and Rubin 1997). Unlike discrepancy variables (Meng 1994), which may depend on unobserved factors (e.g., missing compliance statuses), statistics must be functions of only the observed data. Calculate T on the observed data to obtain T^{obs}.

for $m : 1$ to M **do**

2. **Impute missing compliance statuses.**

   Impute the missing compliance statuses, drawing once from their posterior predictive distribution according to a compliance model that assumes the null hypothesis.

3. **Impute missing potential outcomes.**

   Impute the missing Y potential outcomes under the sharp null hypothesis. Under the typical sharp null hypothesis of zero treatment effect, the missing potential outcome for unit i is imputed exactly as Y_i^{obs}.

4. **Draw a random hypothetical assignment.**

   Draw a random hypothetical assignment vector according to the assignment mechanism.

5. **Re-observe the data.**

   Treating the imputed compliance statuses, imputed potential outcomes, and hypothetical assignment vector from Steps 2–4 as true, create a corresponding hypothetical observed dataset by masking the potential outcomes and compliance statuses that would not have been observed under the hypothetical assignment.

6. **Calculate the test statistic on these data.**

   Calculate T on the hypothetical observed data to obtain T^{hyp}. Record whether this statistic is at least as extreme as T^{obs}.

end for

7. **Calculate the posterior predictive p-value.**

   The posterior predictive p-value for the null hypothesis with respect to T equals the proportion of the M imputation-randomization sets for which T^{hyp} is as extreme as or more extreme than T^{obs}.

Rubin (1998) discusses several commonly used statistics for evaluating complier causal effects, only some of which tend to estimate CACE and thus provide suitable power against appropriate alternative hypotheses. As is commonly done in non-compliance literature, we assume the exclusion restriction (i.e., we assume that treatment assignment has no effect on the outcomes of never-takers and always-takers) for test statistic calculations throughout this paper. Such an assumption is not necessary and does not affect the validity of the randomization test, but it does facilitate more precise estimation of CACE when true (see Imbens and Rubin 1997) and is often reasonable.

The imputation in Step 2 is performed probabilistically, using the missing statuses' null posterior predictive distribution, given X, Z, D^{obs}, and Y^{obs}. (Some test statistics, such as the

posterior median of CACE, may be computed by multiply imputing the missing compliance statuses. This would be a separate imputation from the one described in Step 2 above. If the test statistic calculation itself involves imputation, such imputation does not need to, and usually does not, assume the null hypothesis.) The repetition of Steps 2–6 is intended to reflect the uncertainty of estimation resulting from the missing compliance statuses; M is a large number (e.g., $10,000$) that controls the Monte Carlo integration error.

Under the null hypothesis, Y is not affected by assignment to or receipt of the active treatment; it is therefore treated like a covariate in the imputation model. Even in the absence of other covariates (X), Y alone may still be successful in stochastically identifying the missing compliance statuses, thus providing tests of CACE with power over ITT tests (see Section 3.2.3). When additional covariates that affect compliance status supplement Y in the imputation model (e.g., in a Bayesian generalized linear model), the compliance identification tends to sharpen, providing CACE tests with greater power.

In settings with one-sided non-compliance, only the compliance statuses of units assigned to the control group are missing. Let ω_c be the super-population proportion of compliers, and let $\boldsymbol{\eta} = (\eta_c, \eta_n)$ be the parameters that govern the outcome distributions of compliers and never-takers, respectively. Note that under the null hypothesis, these are only two outcome distributions; units within a compliance stratum have the same outcome distributions, regardless of their treatment assignment. The posterior predictive distribution of the missing compliance statuses can be obtained using a two-step data augmentation algorithm (Tanner and Wong 1987). Using the current (or initial, if starting the algorithm) values of the parameters, the missing compliance statuses are drawn according to Bayes' rule:

$$P(C_i = c | Y_i^{\text{obs}}, X_i, Z_i = 0, D_i^{\text{obs}} = 0, \omega_c, \boldsymbol{\eta}) = \frac{\omega_c g_c(Y_i^{\text{obs}}; \eta_c)}{\omega_c g_c(Y_i^{\text{obs}}; \eta_c) + (1 - \omega_c) g_n(Y_i^{\text{obs}}; \eta_n)}, \quad (3.1)$$

where $g_c(y; \eta_c)$ and $g_n(y; \eta_n)$ are the outcome probabilities (or densities) of y for compliers and never-takers, respectively. Once the missing compliance statuses are drawn, new parameter values are drawn from their compliance-complete-data posterior distributions. These two steps are alternated until distributional convergence. After convergence, the draws of the

missing compliance statuses can be treated as posterior predictive imputations. Obtaining posterior draws of parameters — and consequently, posterior predictive draws of the missing compliance statuses — may be more straightforward if models are conjugate, e.g., Beta-Binomial or Dirichlet-Multinomial models (see Section 3.2.3).

For each imputation of the missing compliance statuses, a randomization test (here involving only one random hypothetical assignment for computational efficiency) is performed in Steps 3–6. Because p-values are defined as conditional probabilities given that the sharp null hypothesis is true, the imputation of Y potential outcomes in Step 3 must occur under this hypothesis. Table 3.3 shows the observed values of the Science table from Table 3.2, with the Y potential outcomes imputed under the sharp null hypothesis of zero treatment effect. For computational efficiency, Step 3 can be performed just once (before the loop) because this imputation is deterministic.

		Assignment	$D(z)$		Compliance status	$Y(z)$	
Unit	X_i	Z_i	$D_i(0)$	$D_i(1)$	C_i	$Y_i(0)$	$Y_i(1)$
1	X_1	0	0	?	?	Y_1^{obs}	(Y_1^{obs})
2	X_2	1	0	1	c	(Y_2^{obs})	Y_2^{obs}
3	X_3	1	0	1	c	(Y_3^{obs})	Y_3^{obs}
4	X_4	1	0	0	nt	(Y_4^{obs})	Y_4^{obs}
...					...		
N	X_N	0	0	?	?	Y_N^{obs}	(Y_N^{obs})

Table 3.3: The observed values of the Science table from Table 3.2, with the missing Y potential outcomes imputed under the sharp null hypothesis of zero treatment effect. Imputed Y potential outcomes are in parentheses.

The random draw of a hypothetical assignment vector in Step 4 depends on the specific assignment mechanism used in the experiment, e.g., complete randomization or block randomization. A seemingly alternative procedure to the one described above switches the order of Steps 2 and 4, such that the hypothetical assignment vector is drawn first, and the missing compliance statuses are imputed second. This alternative procedure, however, is exactly equivalent to the one described above because the imputation of the missing compliance statuses under the null hypothesis is influenced by Z only through C^{obs}. Because C^{obs}

is fixed by the actual observed data, reversing the order of Steps 2 and 4 does not affect the overall inferential procedure. Intuitively, we can consider the posterior predictive p-value as a double integral over the missing compliance statuses and the randomization; switching the order of integration does not affect the result.

3.2.3 Illustrative simulations with non-compliance

Consider this modified example from Rubin (1998): suppose a completely randomized double-blind experiment is conducted to investigate the effect of a new drug (provided in addition to standard care) versus standard care alone on Y, which measures the severity of patients' heart attacks in the year following treatment. Y is ordinal, taking on values of 0, 1, and 2 (no, mild, and severe attacks, respectively). We assume that all of the patients survive through the year. We also assume one-sided non-compliance, so our experiment has two groups of patients: compliers and never-takers.

In our simulation, we randomly select $N = 1000$ units from a super-population of 10% compliers and 90% never-takers; the compliers tend to be healthier than the never-takers. We randomly assign $N/2 = 500$ units to control and $N/2$ units to active treatment, observing only the compliance statuses of units assigned to active treatment. For each unit, we generate an observed Multinomial outcome, Y_i^{obs}, according to the specified treatment effect hypothesis. Simulation details are provided in Appendix A.1.

Using the simulated observed data, we calculate two test statistics: (i) the ITT statistic, and (ii) the MLE of CACE under the exclusion restriction. We then calculate randomization-based posterior predictive p-values for both test statistics, as described in Section 3.2.2, under the null hypothesis of zero treatment effect. (For the multiple imputation of the missing compliance statuses, we place conjugate Beta$(1, 1)$ priors on the parameters governing the complier and never-taker outcome distributions.) To evaluate the frequency characteristics of the posterior predictive p-values, we run 1,000 replications of the data simulation and p-value procedures. Under the null hypothesis, p-values for the two statistics both appear

valid in terms of type I error; their empirical distributions are approximately uniform. At the $\alpha = .05$ level, tests on ITT and CACE reject the null hypothesis in 4.5% and 4.1% of simulations, respectively. Under the alternative hypothesis, tests based on the CACE are more powerful (see Figure 3.1), with tests on ITT and CACE rejecting the null hypothesis in 16.7% and 25.2% of simulations, respectively, at $\alpha = .05$. In a general setting, the magnitude of the power gain from CACE depends on the proportion of compliers, the magnitude of the treatment effect, and the α level.

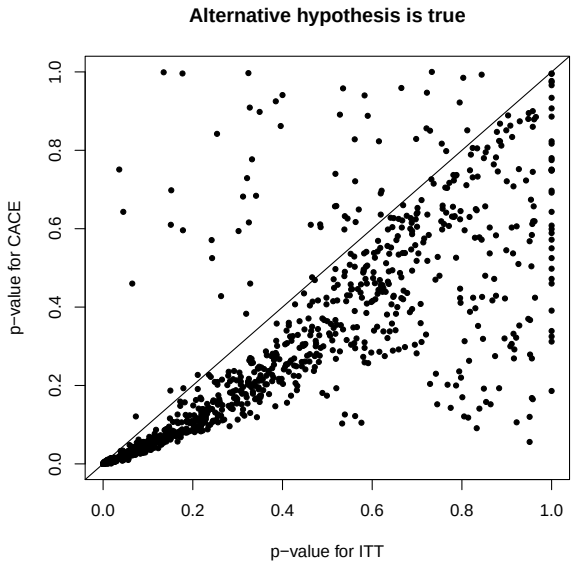

Figure 3.1: Joint distribution of 1,000 posterior predictive p-values for ITT and CACE estimates under the alternative hypothesis. Tests for CACE are more powerful because p-values for CACE tend to be lower.

3.3 Experiments with Multiple Testing

3.3.1 Randomization-based multiple comparisons adjustments

Suppose we have data from a randomized experiment with J estimands and are interested in testing whether the active treatment has any non-null effects. The desire for J estimands may result, for example, from multiple outcomes per unit or from multiple, potentially overlapping

subgroups of units. Brown and Fears (1981) and Westfall and Young (1989) first proposed permutation-based multiple comparisons adjustments, with the latter showing that such adjustments outperform traditional (e.g., Bonferroni) adjustments in terms of power. They did not, however, explicitly motivate their methods using randomized assignment mechanisms and joint randomization distributions. Furthermore, they assumed specific models that facilitated the calculation of nominal (unadjusted) p-values and implicitly assumed completely randomized assignments throughout.

Here we extend their ideas to a fully randomization-based procedure for multiple comparisons adjustments. In contrast to the aforementioned work, our procedure is connected to — and directly motivated by — the actual randomized assignment mechanism used in the experiment; in addition, both the nominal and adjusted p-values in our procedure are randomization-based, so we do not require any assumptions about the underlying distribution of the data. We calculate fully randomization-based adjusted p-values as follows:

1. **Choose test statistics and calculate their observed values.**

   Choose test statistics, $(T_1, \ldots, T_J)$, and calculate $(T_1^{\text{obs}}, \ldots, T_J^{\text{obs}})$ on the observed data.

2. **Impute missing potential outcomes.**

   Impute the missing potential outcomes under the sharp null hypothesis.

3. **Calculate nominal p-values for the observed test statistics.**

   For $j = 1, \ldots, J$, calculate the randomization-based p-value for T_j^{obs} by repeatedly (i) drawing a random hypothetical assignment vector according to the assignment mechanism, and (ii) calculating the test statistic, T_j^{hyp}, for the corresponding hypothetical observed data. The nominal, marginal randomization-based p-value for T_j^{obs} $(j = 1, \ldots, J)$ equals the proportion of T_j^{hyp} values that are as extreme as or more extreme than T_j^{obs}.

for $m' : 1$ to M' **do**

4. **Calculate nominal (marginal) p-values for hypothetical test statistics.**

   Draw a random hypothetical treatment assignment according to the assignment mechanism and calculate test statistics $(T_1^{\text{hyp}}, \ldots, T_J^{\text{hyp}})$ on the corresponding hypothetical observed data. Calculate a nominal randomization-based p-value for each T_j^{hyp} and record the minimum of the p-values.

end for

5. **Obtain the joint randomization distribution of the nominal p-values.**

   For large M', the repetitions of Step 4 appropriately capture the joint randomization distribution of the test statistics and thus, of the nominal p-values.

6. **Calculate adjusted p-values for the observed test statistics.**

   The adjusted p-value (Westfall and Young 1989) for T_j^{obs} $(j = 1, \ldots, J)$ equals the proportion of hypothetical observed datasets for which the minimum of the J nominal p-values for $(T_1^{\text{hyp}}, \ldots, T_J^{\text{hyp}})$ is less than or equal to the nominal p-value for T_j^{obs}.

For computational efficiency, the hypothetical randomizations and statistic calculations in Step 3 can be recycled for Steps 4–5: Steps 4–5 essentially represent a translation, i.e., re-scaling, of hypothetical test statistics — which may have different scales — into hypothetical p-values, which share a common 0–1 scale. The procedure described above results in individual adjusted p-values that are corrected for the FWER but are also directly interpretable on their own.

Equivalently, to determine α-level significance, we can compare each nominal p-value to the familywise α-level cutoff: the α-th quantile of the minimums recorded in Step 4. The probability that no type I errors are made (i.e., that we fail to reject all J tests under the null hypothesis) is equivalent to the probability that all J observed marginal p-values are above the cutoff. This equals the probability that the minimum of the J observed p-values is above the cutoff, which is $1 - \alpha$ by construction. Thus, the probability of at least one type I error — the FWER — is α, as desired.

Randomization-adjusted p-values are more powerful than traditional Bonferroni-adjusted p-values, especially when the correlations among the J test statistics are high, as shown by the simulations below. Intuitively, suppose the null hypothesis is true and that we have a large number of uncorrelated test statistics; the probability of at least one type I error is quite high because of the number of tests being conducted. Now suppose instead that those test statistics are highly correlated; the probability of at least one type I error is reduced because the tests' type I errors are likely to occur simultaneously, i.e., for the same random assignments. In fact, if the test statistics are perfectly correlated, there is essentially only one test being conducted, so no multiple comparisons adjustment is needed. Bonferroni adjustments in all of these settings are the same, simply counting the number of p-values being examined. In contrast, by utilizing the joint distribution of the nominal p-values, the randomization-based adjustments account for the correlations among test statistics and are less conservative.

3.3.2 Illustrative simulations with multiple testing

We follow the experimental setup from Section 3.2.3, modified to include multiple outcomes but without non-compliance. Suppose that researchers now want to investigate the effect of the new drug on three outcomes: $Y_{\cdot 1}, Y_{\cdot 2}$, and $Y_{\cdot 3}$ (with the first subscript denoting the participant), which measure the severity of heart attacks (defined as before) in the first, second, and third year after treatment, respectively. We assume that all of the patients survive through the third year, and we would like to see whether the drug has an effect on heart attack severity at any of the three time points.

To evaluate the frequency characteristics of the adjusted randomization-based p-values, we simulate 1,000 datasets under both null and alternative hypotheses according to each of three outcome correlation structures: zero, partial (approximately 0.5), and perfect correlation. The specific data generation processes are found in in Appendices A.2 and B. The correlations among $Y_{i1}(z)$, $Y_{i2}(z)$, and $Y_{i3}(z)$ $(z = 0, 1)$ are important; however, for a fixed

j, the correlation between $Y_{ij}(0)$ and $Y_{ij}(1)$ is inconsequential to the simulation because we only ever observe one of the potential outcomes.

For each simulated dataset, we calculate fully randomization-based adjusted p-values and decide whether or not to reject the null hypothesis of zero treatment effects across the three time periods at $\alpha = .05$. For comparison, we also decide whether or not to reject the null hypothesis using Bonferroni-adjusted p-values. Simulation results under both null and alternative hypotheses are shown in Table 3.4. Without sacrificing validity under the null hypothesis, the randomization-based adjustment displays greater power than the Bonferroni adjustment under the alternative hypothesis, particularly for scenarios with high correlations among outcomes.

	Rejection Rate at $\alpha = .05$			
	Null is true		Alternative is true	
	Bonferroni	Randomization-Based	Bonferroni	Randomization-Based
Zero correlation	.042	.046	.908	.919
Partial correlation	.045	.053	.787	.811
Perfect correlation	.024	.045	.557	.720

Table 3.4: Proportions of multiple testing simulations in which the null hypothesis was rejected, under various data generation processes. Based on 1,000 replications.

3.4 Experiments with Both Non-compliance and Multiple Testing

It is natural to merge the analysis methods presented in Sections 3.2 and 3.3 — both of which use the randomized assignment mechanism to aid inference — for experiments involving both non-compliance and multiple testing. The results are valid familywise tests that are doubly more powerful: more powerful than both those based on standard ITT statistics and those using traditional multiple comparison adjustments.

Suppose again that we have data from a randomized experiment with J estimands and that we are interested in testing whether the active treatment has any non-null effects.

However, not all units comply to their treatment assignments; assume for simplicity that non-compliance is one-sided. In Section 3.2, Table 3.2 displays the observed values of a Science table with two Y potential outcomes — one observed and one missing — for each unit. Here, Table 3.5 shows the corresponding observed values of a Science table with multiple estimands resulting from $J = 3$ outcomes of interest. Each unit has six potential outcomes, only three of which are observed; the other three are missing. Within unit i, we observe the same member of $(Y_{ij}(0), Y_{ij}(1))$ for each outcome j, e.g., if we observe $Y_{i1}(1)$, then we also observe $Y_{i2}(1)$ and $Y_{i3}(1)$.

		Assignment	$D(z)$		Compliance status	$Y_{\cdot 1}(z)$		$Y_{\cdot 2}(z)$		$Y_{\cdot 3}(z)$	
Unit	X_i	Z_i	$D_i(0)$	$D_i(1)$	C_i	$Y_{i1}(0)$	$Y_{i1}(1)$	$Y_{i2}(0)$	$Y_{i2}(1)$	$Y_{i3}(0)$	$Y_{i3}(1)$
1	X_1	0	0	?	?	Y_{11}^{obs}	?	Y_{12}^{obs}	?	Y_{13}^{obs}	?
2	X_2	1	0	1	c	?	Y_{21}^{obs}	?	Y_{22}^{obs}	?	Y_{23}^{obs}
3	X_3	1	0	1	c	?	Y_{31}^{obs}	?	Y_{32}^{obs}	?	Y_{33}^{obs}
4	X_4	1	0	0	nt	?	Y_{41}^{obs}	?	Y_{42}^{obs}	?	Y_{43}^{obs}
...			...					...			
N	X_N	0	0	?	?	Y_{N1}^{obs}	?	Y_{N2}^{obs}	?	Y_{N3}^{obs}	?

Table 3.5: Observed and unobserved values of the Science table from Table 3.2, now with three outcomes of interest. Missing (unobserved) data are denoted by question marks.

In experiments with non-compliance and multiple testing, obtaining valid and doubly more powerful familywise tests involves (i) calculating (nominal) posterior predictive p-values for CACE according to the procedure in Section 3.2, and (ii) calculating adjusted posterior predictive p-values using the joint randomization distribution of the nominal p-values, according to the procedure in Section 3.3. Intuitively, this combined method of analysis is preferable because Steps (i) and (ii) provide power gains through distinct and unrelated mechanisms, and neither sacrifices validity in terms of type I error. For the J estimands, we expect each individual (nominal) CACE p-value to be more powerful than its ITT counterpart based on the arguments in Section 3.2. Furthermore, given a set of J nominal p-values, we expect randomization-adjusted p-values using the nominal p-values' joint randomization distribution to be more powerful than Bonferroni-adjusted p-values, as argued in Section 3.3. Naturally, adjusting more powerful nominal p-values in a more powerful manner results in doubly more powerful adjusted p-values. The full procedure is detailed below:

1. **Choose test statistics and calculate their observed values.**

   Choose test statistics, $(T_1, \ldots, T_J)$, and calculate $(T_1^{\text{obs}}, \ldots, T_J^{\text{obs}})$ on the actual observed data.

for $i : 1$ to M **do**

2. **Impute missing compliance statuses.**

   Impute the missing compliance statuses, drawing once from their posterior predictive distribution according to a compliance model that assumes the null hypothesis.

3. **Impute missing potential outcomes.**

   Impute all of the missing $(Y_1, \ldots, Y_J)$ potential outcomes under the sharp null hypothesis.

4. **Draw a random hypothetical assignment.**

   Draw a random hypothetical assignment vector according to the assignment mechanism.

5. **Re-observe the data.**

   Treating the imputed compliance statuses and potential outcomes and the hypothetical assignment vector as true, create a corresponding hypothetical observed dataset by masking the potential outcomes and compliance statuses that would not have been observed under the hypothetical assignment.

6. **Calculate test statistics on the hypothetical observed data.**

   Calculate $(T_1, \ldots, T_J)$ on the hypothetical observed data to obtain $(T_1^{\text{hyp}}, \ldots, T_J^{\text{hyp}})$. For $j = 1, \ldots, J$, record whether T_j^{hyp} is at least as extreme as T_j^{obs}.

end for

7. **Calculate nominal (marginal) posterior predictive p-values for the observed test statistics.**

   For $j = 1, \ldots, J$, the nominal (marginal) posterior predictive p-value for the null hy-

pothesis with respect to the test statistic T_j equals the proportion of the M imputation-randomization sets created by Steps 2–6 for which T_j^{hyp} is as extreme as or more extreme than T_j^{obs}.

8. **Calculate nominal posterior predictive p-values for hypothetical test statistics and obtain the joint randomization distribution of the nominal posterior predictive p-values.**

   For each of the M imputation-randomization sets, translate the hypothetical test statistics $(T_1^{\text{hyp}}, \ldots, T_J^{\text{hyp}})$ into hypothetical nominal posterior predictive p-values using proportions similar to the one described in Step 7. This step is a computationally efficient way of obtaining the joing distribution of hypothetical test statistics on a common p-value scale, analogous to Steps 4–5 from the procedure in Section 3.3. Record the minimum of each set of nominal p-values.

9. **Calculate adjusted posterior predictive p-values for the observed test statistics.**

   The adjusted posterior predictive p-value for T_j^{obs} $(j = 1, \ldots, J)$ equals the proportion of the M imputation-randomization sets for which the minimum of the J nominal posterior predictive p-values for $(T_1^{\text{hyp}}, \ldots, T_J^{\text{hyp}})$ is less than or equal to the nominal (marginal) posterior predictive p-value for T_j^{obs}.

Under the null hypothesis, the outcomes $Y_{\cdot 1}, \ldots, Y_{\cdot J}$ inform the multiple imputation of the missing compliance statuses. Posterior predictive imputations of the missing compliance statuses can be generated using a data augmentation algorithm similar to the one described in Section 3.2, with Equation 3.1 modified to use the joint set of J observed outcomes.

3.4.1 Illustrative simulations with both non-compliance and multiple testing

Again consider the heart treatment example from Sections 3.2.3 and 3.3.2: we would like to see whether the active treatment has an effect on heart attack severity at any of the three time points after treatment. In these simulations, we assume one-sided non-compliance, with $N = 1000$ units randomly sampled from a super-population of 10% compliers and 90% never-takers. The data generation processes are described in Appendices A.3 and B.

For each simulated dataset, four familywise tests are conducted. Two of the tests use the ITT test statistic, one with the Bonferroni correction and the other with the randomization-based multiple comparisons adjustment. The other two tests use the MLE of CACE (under the exclusion restriction) as the test statistic, one with the Bonferroni correction and the other with the randomization-based adjustment. Table 3.6 displays proportions of simulations in which the null hypothesis was rejected, based on 500 replications.

	Rejection Rate at $\alpha = .05$			
	ITT		CACE	
Null is true	Bonferroni	Randomization-Based	Bonferroni	Randomization-Based
Zero correlation	.036	.042	.022	.024
Partial correlation	.022	.036	.018	.030
Perfect correlation	.016	.056	.008	.040
	ITT		CACE	
Alternative is true	Bonferroni	Randomization-Based	Bonferroni	Randomization-Based
Zero correlation	.198	.222	.240	.260
Partial correlation	.114	.148	.178	.206
Perfect correlation	.092	.184	.142	.256

Table 3.6: Proportions of simulations in which the null hypothesis was rejected, under various data generation processes. Based on 500 replications.

Under the null hypothesis, all four familywise tests appear valid in terms of type I error. The randomization-based tests have the rejection rates closest to the nominal rejection rates. As expected, the Bonferroni-adjusted tests are conservative, especially when correlation among outcomes is high. In such settings, there are, in a sense, fewer possible

effects to detect, and randomization-adjusted rejection rates are much higher relative to their Bonferroni-adjusted counterparts.

Under alternative hypotheses, the CACE tests generally have higher power, i.e., higher rejection rates, than the ITT tests. In addition, the randomization-based tests outperform their Bonferroni counterparts, especially when correlation among outcomes is high. In our simulations, CACE tests with randomization-based multiple comparisons adjustments have 30% to 175% higher relative power than traditional Bonferroni ITT tests. In a particular experimental setting, the magnitude of the power gain from the combined analysis method depends on the compliance rate, the magnitude of the treatment effect, the α level, and the correlation of the multiple test statistics.

3.5 The National Job Training Partnership Act Study

Title II of the United States Job Training Partnership Act (JTPA) of 1982 funded employment training programs for economically disadvantaged residents (Bloom et al. 1997; Abadie et al. 2002). To evaluate the effectiveness of those training programs, the National JTPA Study conducted a randomized experiment through 16 local administration areas involving a total of around 20,000 participants who applied for JTPA services from November 1987 to September 1989 (W.E. Upjohn Institute for Employment Research 2013). Treatment group participants were eligible to receive JTPA services, while control group participants were ineligible to receive JTPA services for 18 months. Not every participant assigned to the treatment group actually enrolled and received JTPA services.

3.5.1 The data

Monthly employment outcomes were recorded for 30 months after assignment through follow-up surveys and administrative records from state unemployment insurance agencies. Researchers were interested in measuring JTPA effects across three time periods represent-

ing various stages of training and employment: months 1–6 (after assignment), the period when most JTPA enrollees were in the program; months 7–18, approximately the first post-program year; and months 19–30, approximately the second post-program year (Bloom et al. 1997).

Bloom et al. (1997)'s original JTPA report evaluates effects on average income but does not explicitly address the large portion of zero-income (i.e., unemployed) participants. Although the report describes effects by subperiod as well as by various participant subgroups, it fails to mention or employ any multiple comparisons adjustments. Here we focus on JTPA's effects on employment status and use gender as our only background covariate; this facilitates standard, non-controversial modeling choices (see Section 3.5.2) and allows us to highlight our methodological contributions rather than discuss the sensitivity of our results to various, possibly complicated modeling decisions. Our methods can be extended to evaluate effects on other outcome variables, such as income and wages, provided that we outline a reasonable imputation model (Zhang et al. 2009).

We would like to evaluate whether JTPA had an effect on employment status for any of the three time periods. Because employment characteristics often differ by gender, we examine JTPA effects for the three time periods by gender, for a total of six gender-time groups. For illustrative purposes, we restrict our study population to adults who had obtained a high school or GED diploma (7,445, or 66.4%, of the 11,204 total adults in the original JTPA study) and assume complete randomization (with an approximate $2:1$ treatment-to-control assignment ratio) of the participants, ignoring the local administration structure because of the limitations of the available data.

Of the 5,009 participants assigned to the treatment group, 3,316 (66.2%) subsequently received JTPA training. Although the study protocol barred participants assigned to the control group from receiving JTPA services for 18 months, 41 (1.7%) of 2,436 adults in the control group did in fact receive services within that time frame. To create a simpler setting with true one-sided non-compliance, we discard these 41 participants (0.6% of the 7,445 total

adults in our study) with the belief that their inclusion would have a negligible influence on the resulting inference.

Given two genders and three time periods, we have six complier-focused estimands in total, each one representing the difference in employment proportions within a particular gender-time group when receiving versus not receiving JTPA services. Two summaries of the observed data are provided in Figure 3.2 and Table 3.7. Figure 3.2 shows observed employment proportions across the six gender-time groups by observed compliance status. Within every group, observed compliers are employed at a higher rate than observed never-takers. Participants with unobserved compliance statuses (i.e., those assigned to control) are a mixture of compliers and never-takers, and tend to be employed at a rate in between the rates for observed compliers and observed never-takers.

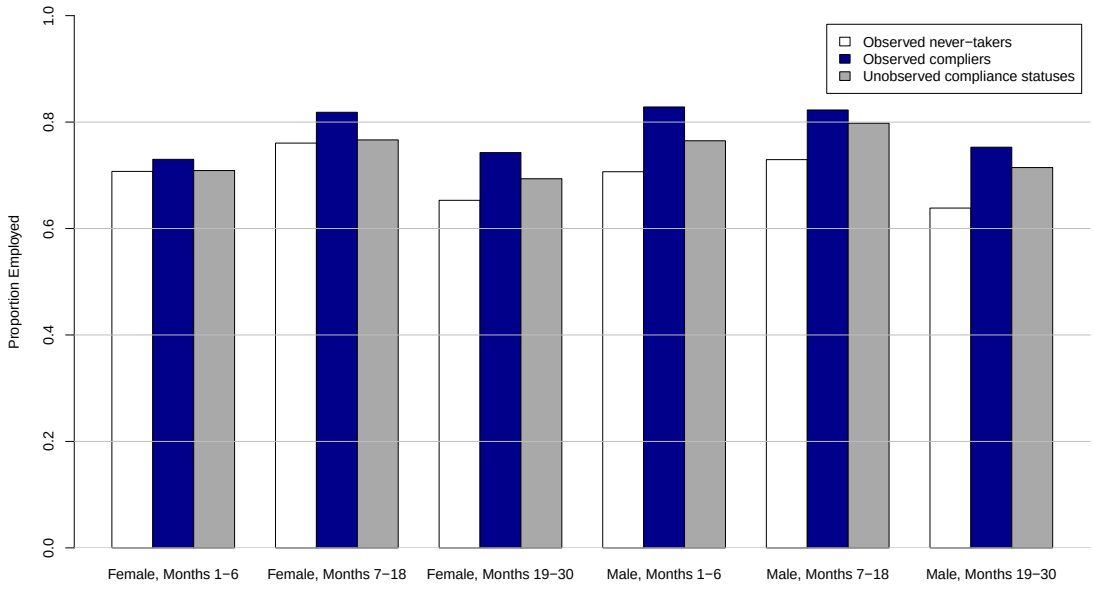

Figure 3.2: Observed employment proportions for JTPA participants by compliance status across the six gender-time groups.

Table 3.7 displays observed employment proportions across the gender-time groups according to both treatment assignment and treatment receipt, with the corresponding compliance compositions. We see that participants who received JTPA services, all of whom are

	Observed Employment Proportions	
	Assigned Control $C_i \in \{c, nt\}, Z_i = 0$	Assigned Treatment $C_i \in \{c, nt\}, Z_i = 1$
Female, Months 1–6	.709	.723
Female, Months 7–18	.767	.800
Female, Months 19–30	.694	.714
Male, Months 1–6	.765	.785
Male, Months 7–18	.798	.789
Male, Months 19–30	.715	.712
	Received Control $(C_i \in \{c, nt\}, Z_i = 0)$ or $(C_i = nt, Z_i = 1)$	Received Treatment $C_i = c, Z_i = 1$
Female, Months 1–6	.708	.730
Female, Months 7–18	.764	.818
Female, Months 19–30	.677	.743
Male, Months 1–6	.740	.828
Male, Months 7–18	.769	.823
Male, Months 19–30	.683	.753

Table 3.7: Observed employment proportions across the six gender-time groups according to both assignment to and receipt of JTPA services.

compliers, tend to be employed at a higher rate than participants who were merely assigned to the treatment group (a mixture of compliers and never-takers), corroborating the findings in Figure 3.2 and suggesting that CACE statistics may lead to more significant estimated effects. In addition, we observe that participants who did not receive JTPA services — including any participants assigned to control as well as the never-takers assigned to JTPA — are employed at a lower rate than just the participants assigned to control. This inequality is intuitive because the observed never-takers are shown in Figure 3.2 to be employed at a lower rate than the assigned control group.

3.5.2 Imputation model for CACE

To test the null hypothesis of zero effects using the CACE statistic specified in Section 3.2, we must specify an imputation model for the missing compliance statuses. Let X_i and $\boldsymbol{Y_i}$ denote the gender and the length-3 vector of employment outcomes (across the three time periods) of participant i. The three elements of $\boldsymbol{Y_i}$ are binary, so there are $2^3 = 8$ possible

values of $\boldsymbol{Y_i}$; we model $\boldsymbol{Y}$ as a Multinomial random variable with eight categories. Let ω_c be the super-population proportion of compliers, and let $\boldsymbol{\eta} = (\eta_{fc}, \eta_{fn}, \eta_{mc}, \eta_{mn})$ be the parameters that govern the outcome distributions of female compliers, female never-takers, male compliers, and male never-takers, respectively. Under the null hypothesis, these are the only four outcome distributions because we disregard treatment assignment. We place a conjugate Beta(1,1) prior on ω_c and independent conjugate Dirichlet($\mathbf{1}$) priors on the four η parameters, where $\mathbf{1}$ is a length-8 vector of 1's.

Conditional on $\boldsymbol{\eta}$ and a participant's gender and compliance status, the natural outcome distribution under the null hypothesis is:

$$\boldsymbol{Y_i}^{\text{obs}} | X_i = x, C_i = q, \boldsymbol{\eta} \sim \text{Multinomial}(1, \eta_{xq}).$$

Note that we do not assume that the three employment outcomes are independent; this model is fully non-parametric for the joint distribution of the three outcomes. The posterior distributions of ω_c and $\boldsymbol{\eta}$ are informed by the outcomes of the participants with observed compliance statuses, i.e., those assigned to active treatment, and remain Beta and Dirichlet, respectively. For each gender x and compliance status q, write the Multinomial probability vector as

$$\eta_{xq} = (\pi_{xq1}, ..., \pi_{xq7}, 1 - \pi_{xq1} - \ldots - \pi_{xq7}).$$

Let

$$g_{xq}(\boldsymbol{y}; \eta_{xq}) = \pi_{xq1}^{I\{\boldsymbol{y}=(0,0,0)\}} \pi_{xq2}^{I\{\boldsymbol{y}=(0,0,1)\}} \ldots (1 - \pi_{xq1} - \ldots - \pi_{xq7})^{I\{\boldsymbol{y}=(1,1,1)\}}$$

denote the probability of outcome $\boldsymbol{y}$ for participants of gender x and compliance status q. Then, given a posterior draw of $(\omega_c, \boldsymbol{\eta})$, the missing compliance statuses are imputed probabilistically according to Bayes' rule:

$$P(C_i = c | \boldsymbol{Y_i^{\text{obs}}}, X_i = x, Z_i = 0, \omega_c, \boldsymbol{\eta}) = \frac{\omega_c g_{xc}(\boldsymbol{Y_i^{\text{obs}}}; \eta_{xc})}{\omega_c g_{xc}(\boldsymbol{Y_i^{\text{obs}}}; \eta_{xc}) + (1 - \omega_c) g_{xn}(\boldsymbol{Y_i^{\text{obs}}}; \eta_{xn})}. \quad (3.2)$$

3.5.3 Results and analysis

The observed values of the ITT and CACE statistics — i.e., the estimated effects of JTPA assignment and of receipt, respectively — are shown in the second column of Table 3.8. As we expect, the observed CACE values have larger magnitudes; the estimated ITT effects are diluted toward zero by the never-takers, who do not receive any treatment benefit. Because $\text{ITT} = \omega_c * \text{CACE} + (1 - \omega_c) * 0$, the estimated ITT effects are diluted by a proportion equal to one minus the compliance rate. Due to the random treatment assignment, we expect the overall compliance rate to be approximately equal to the compliance rate observed in the treatment group (66.2%).

			Adjusted p-values	
ITT	Estimated Effect	Nominal p-value	Bonferroni	Randomization
Female, Months 1–6	.014	.351	1.000	.895
Female, Months 7–18	.020	.199	1.000	.685
Female, Months 19–30	.033	.014	.085	.077
Male, Months 1–6	-.008	.582	1.000	.991
Male, Months 7–18	.020	.175	1.000	.636
Male, Months 19–30	-.003	.874	1.000	1.000
			Adjusted p-values	
CACE	Estimated Effect	Nominal p-value	Bonferroni	Randomization
Female, Months 1–6	.020	.130	.778	.302
Female, Months 7–18	.034	.009	.055	.026
Female, Months 19–30	.049	.0002	.001	.001
Male, Months 1–6	-.010	.462	1.000	.804
Male, Months 7–18	.028	.028	.169	.076
Male, Months 19–30	-.001	.967	1.000	1.000

Table 3.8: Observed values, nominal p-values, and Bonferroni- and randomization-adjusted p-values for the six JTPA gender-time groups. Nominal p-values are obtained through randomization tests using 10,000 randomizations.

Using randomization tests and the methods described in Section 3.4, we obtain one set of nominal ITT p-values and a second set of nominal CACE p-values, listed in the third column of Table 3.8. Each set contains six p-values, one for each gender-time group. We also apply Bonferroni and randomization adjustments to both sets of nominal p-values, resulting in four total sets of adjusted p-values, listed in the rightmost columns of Table 3.8.

The nominal ITT p-value for the "Female, Months 19–30" group indicates statistical significance at the $\alpha = .05$ level. However, after adjusting for multiple comparisons, neither the Bonferroni- nor randomization-adjusted ITT p-values for this group meets the .05 threshold. Across the six gender-time groups, the randomization-adjusted p-values tend to be smaller than their Bonferroni-adjusted counterparts; the adjusted p-values are tempered less when controlling the FWER via the statistics' joint randomization distribution because of the correlations among the six nominal p-values.

Overall, the CACE p-values are smaller — more sensitive to complier-only effects — than the ITT p-values. In particular, the CACE p-values for the "Female, Months 7–18" and "Female, Months 19–30" groups indicate a much greater level of significance for the estimated effects of JTPA on employment. Applying randomization-based instead of Bonferroni adjustments to the CACE p-values further increases the indicated significance of these estimated effects. The small randomization-adjusted CACE p-values for these groups suggest that either an event has occurred that is *a priori* rare under the sharp null hypothesis of zero effects, or the sharp null hypothesis is not true — *receipt* of JTPA services did have an effect on the employment statuses of females with high school or GED diplomas in their first and second post-program years. The corresponding ITT p-values, although smallest among the six groups, are larger and do not have sufficient power to detect an effect on employment status for any of the gender-time groups.

This increase in power is general. We observe similar p-value trends when comparing our methods to ITT and Bonferroni analyses on JTPA data without the high school/GED diploma restriction as well as on other JTPA subgroups analyzed in Bloom et al. (1997).

3.6 Conclusion

We have detailed a randomization-based procedure for analyzing experimental data in the presence of both non-compliance and multiple testing that is more powerful than traditional

ITT and Bonferroni analyses. As shown through simulations and analyses of the National JTPA Study data, a combined randomization-based procedure can be doubly advantageous, offering gains in power from both perspectives.

A number of other multiple comparisons procedures aim to address the false discovery rate (FDR) (Benjamini and Hochberg 1995), rather than the FWER. These two error metrics are conceptually different. We focus on the FWER here; the choice of metric depends on the particular research setting and goals.

Appendix A

Marginal Distributions for Chapter 3 Simulations

A.1 Non-compliance

For unit $i = 1, \ldots, N$, the control potential outcomes for compliers and never-takers have the following marginal distributions:

$$Y_i(0)|C_i = c \sim \text{Multinomial}(.45, .45, .10); \tag{A.1}$$

$$Y_i(0)|C_i = nt \sim \text{Multinomial}(.02, .02, .96). \tag{A.2}$$

Under the null hypothesis, $Y_i(1)$ has the same marginal distribution as $Y_i(0)$ regardless of compliance status. Under the alternative hypothesis, the complier treatment potential outcomes follow:

$$Y_i(1)|C_i = c \sim \text{Multinomial}(.80, .10, .10), \tag{A.3}$$

while the never-taker treatment potential outcomes follow Equation A.2.

A.2 Multiple testing

For unit $i = 1, \ldots, N$ and outcome $j = 1, 2, 3$, the control potential outcomes marginally follow:

$$Y_{ij}(0) \sim \text{Multinomial}(.45, .45, .10). \tag{A.4}$$

Under the null hypothesis, $Y_{ij}(1)$ has the same marginal distribution as $Y_{ij}(0)$. Under the alternative hypotheses, the treatment potential outcomes have the following marginal distribution:

$$Y_{ij}(1) \sim \text{Multinomial}(.50, .45, .05). \tag{A.5}$$

A.3 Non-compliance and multiple testing

The potential outcomes follow the marginal distributions described in Appendix A.1.

Appendix B

Correlation Structure Generation for Chapter 3 Simulations

To simulate correlation structures among multiple outcomes, we use the following processes utilizing the marginal distributions described in Appendix A. For units $i = 1, \ldots, N$ and treatment assignment $z = 0, 1$,

- Zero correlation: all $Y_{ij}(z)$ $(j = 1, 2, 3)$ are drawn independently according to their marginal distributions.
- Partial correlation: $Y_{i1}(z)$ is drawn according to its marginal distribution. With probability 1/2, $Y_{i2}(z)$ is set equal to the drawn value of $Y_{i1}(z)$; otherwise, $Y_{i2}(z)$ is drawn independently according to its marginal distribution. $Y_{i3}(z)$ is set equal to $Y_{i1}(z)$ with probability 1/3, set equal to $Y_{i2}(z)$ with probability 1/3, or drawn independently according to its marginal distribution.
- Perfect correlation: $Y_{i1}(z)$ is drawn according to its marginal distribution. Then both $Y_{i2}(z)$ and $Y_{i3}(z)$ are set equal to the drawn value of $Y_{i1}(z)$.

Bibliography

Abadie, A., Angrist, J., and Imbens, G. (2002), "Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings," *Econometrica*, 70, 91–117.

A.D.A.M., Inc. (2013), "Idiopathic pulmonary fibrosis," A.D.A.M. Medical Library [online]. Available at `http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001134/`.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of causal effects using instrumental variables," *Journal of the American Statistical Association*, 91, 444–455.

Assmann, S. F., Pocock, S. J., Enos, L. E., and Kasten, L. E. (2000), "Subgroup analysis and other (mis) uses of baseline data in clinical trials," *The Lancet*, 355, 1064–1069.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 289–300.

Bloom, H. S., Orr, L. L., Bell, S. H., Cave, G., Doolittle, F., Lin, W., and Bos, J. M. (1997), "The benefits and costs of JTPA Title II-A programs: Key findings from the National Job Training Partnership Act Study," *Journal of Human Resources*, 32, 549–576.

Brown, C. C. and Fears, T. R. (1981), "Exact significance levels for multiple binomial testing with application to carcinogenicity screens," *Biometrics*, 37, 763–774.

Brown, D. (September 23, 2013), "The press-release conviction of a biotech CEO and its impact on scientific research," The Washington Post [online]. Available at `http://www.washingtonpost.com/national/health-science/the-press-release-crime-of-a-biotech-ceo-and-its-impact-on-scientific-`

research/2013/09/23/9b4a1a32-007a-11e3-9a3e-916de805f65d_story.html.

Cabin, R. J. and Mitchell, R. J. (2000), "To Bonferroni or not to Bonferroni: when and how are the questions," *Bulletin of the Ecological Society of America*, 81, 246–248.

Cleveland Clinic (2013), "Idiopathic pulmonary fibrosis," Cleveland Clinic [online]. Available at http://my.clevelandclinic.org/disorders/pulmonary_fibrosis/hic_idiopathic_pulmonary_fibrosis.aspx.

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd, 1st ed.

— (1935), *The design of experiments*, Oxford: Oliver & Boyd.

Frangakis, C. E. and Rubin, D. B. (2002), "Principal stratification in causal inference," *Biometrics*, 58, 21–29.

Good, P. I. (2005), *Permutation, parametric and bootstrap tests of hypotheses*, vol. 3, Springer.

Imbens, G. W. and Rubin, D. B. (1997), "Bayesian inference for causal effects in randomized experiments with noncompliance," *The Annals of Statistics*, 25, 305–327.

InterMune (2012), "InterMune announces phase III data demonstrating survival benefit of Actimmune in IPF," U.S. Securities and Exchange Commission [online]. Available at http://www.sec.gov/Archives/edgar/data/1087432/000091205702033878/a2088367zex-99_1.htm.

— (2013), "Idiopathic pulmonary fibrosis," InterMune [online]. Available at http://www.intermune.com/idiopathic_pulmonary_fibrosis.

King Jr, T. E., Albera, C., Bradford, W. Z., Costabel, U., Hormel, P., Lancaster, L., Noble, P. W., Sahn, S. A., Szwarcberg, J., Thomeer, M., et al. (2009), "Effect of interferon gamma-1b on survival in patients with idiopathic pulmonary fibrosis (INSPIRE): a multicentre, randomised, placebo-controlled trial," *The Lancet*, 374, 222–228.

Lee, J. J. and Dasgupta, T. (2013–2015), *randomizationInference: Flexible Randomization-Based Inference*, R package version 1.0.3. Available at `http://CRAN.R-project.org/package=randomizationInference`.

Lee, J. J., Dasgupta, T., and Rubin, D. B. (submitted), "Randomization-based inference for industrial experiments," *Technometrics*.

Lee, J. J., Miratrix, L., and Pillai, N. S. (submitted), "More powerful multiple testing in randomized experiments with non-compliance," *Statistica Sinica*.

Lee, J. J. and Rubin, D. B. (accepted, 2015a), "Evaluating the validity of post-hoc subgroup inferences: A case study," *The American Statistician*.

— (accepted, 2015b), "Valid randomization-based p-values for partially post-hoc subgroup analyses," *Statistics in Medicine*.

Meng, X.-L. (1994), "Posterior predictive p-values," *The Annals of Statistics*, 22, 1142–1160.

Miller, R. G. (1981), *Simultaneous statistical inference*, Springer New York, 2nd ed.

Morgan, K. L. and Rubin, D. B. (2012), "Rerandomization to improve covariate balance in experiments," *The Annals of Statistics*, 40, 1263–1282.

Nakagawa, S. (2004), "A farewell to Bonferroni: the problems of low statistical power and publication bias," *Behavioral Ecology*, 15, 1044–1045.

Neyman, J. (1923), "On the application of probability theory to agricultural experiments. Essay on principles. Section 9," *Statistical Science*, 5, 465–472, translated by Dabrowska, DM and Speed, TP (1990).

Peck, L. R. (2003), "Subgroup analysis in social experiments: Measuring program impacts based on post-treatment choice," *American Journal of Evaluation*, 24, 157–187.

Perneger, T. V. (1998), "What's wrong with Bonferroni adjustments," *British Medical Jour-*

nal, 316, 1236–1238.

Q-Med, AB (2009), "Durolane Knee Premarket Approval Application (P060013), Sponsor Executive Summary," Presented to the U.S. Food and Drug Administration.

Raghu, G., Brown, K. K., Bradford, W. Z., Starko, K., Noble, P. W., Schwartz, D. A., and King Jr, T. E. (2004), "A placebo-controlled trial of interferon gamma-1b in patients with idiopathic pulmonary fibrosis," *New England Journal of Medicine*, 350, 125–133.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Rothman, K. J. (1990), "No adjustments are needed for multiple comparisons." *Epidemiology*, 1, 43–46.

Rothwell, P. M. (2005), "Subgroup analysis in randomised controlled trials: importance, indications, and interpretation," *The Lancet*, 365, 176–186.

Rubin, D. B. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, 66, 688–701.

— (1980), "Comment on Randomization analysis of experimental data: the Fisher randomization test," *Journal of the American Statistical Association*, 75, 591–593.

— (1984), "Bayesianly justifiable and relevant frequency calculations for the applied statistician," *The Annals of Statistics*, 12, 1151–1172.

— (1986), "Comment: Which ifs have causal answers," *Journal of the American Statistical Association*, 81, 961–962.

— (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, Inc.

— (1998), "More powerful randomization-based p-values in double-blind trials with noncompliance," *Statistics in Medicine*, 17, 371–385.

— (2005), "Causal inference using potential outcomes," *Journal of the American Statistical Association*, 100, 322–331.

Stretch, B. J., Waldinger, K. F., and Gordus, A. (2010), "United States' sentencing memorandum," Case No. CR 08-164 MHP. Unpublished legal document.

Tanner, M. A. and Wong, W. H. (1987), "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, 82, 528–540.

United States v. Harkonen (2013), 510 Fed. Appx. 633 (9th Cir.).

Tukey, J. W. (1991), "The philosophy of multiple comparisons," *Statistical Science*, 6, 100–116.

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., and Drazen, J. M. (2007), "Statistics in medicine - reporting of subgroup analyses in clinical trials," *New England Journal of Medicine*, 357, 2189–2194.

W.E. Upjohn Institute for Employment Research (2013), "The National JTPA Study," Public Use Data and Data Summary.

Westfall, P. H. and Young, S. S. (1989), "p value adjustments for multiple tests in multivariate binomial models," *Journal of the American Statistical Association*, 84, 780–786.

Zhang, J. L., Rubin, D. B., and Mealli, F. (2009), "Likelihood-based analysis of causal effects of job-training programs using principal stratification," *Journal of the American Statistical Association*, 104, 166–176.